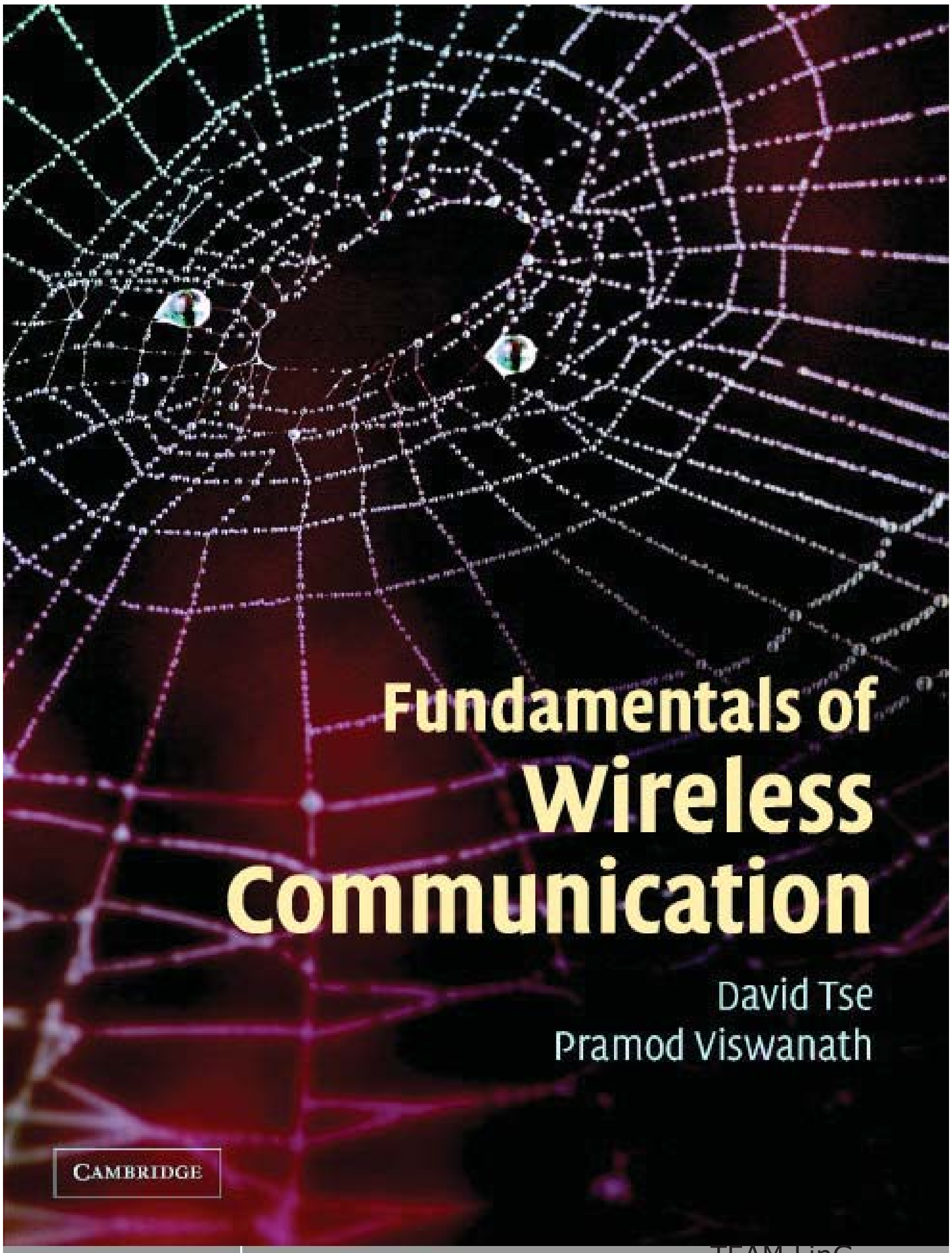


Fundamentals of Wireless Communication

David Tse
Pramod Viswanath

CAMBRIDGE

CAMBRIDGE
more information - www.cambridge.org/9780521845274

Fundamentals of Wireless Communication

The past decade has seen many advances in physical-layer wireless communication theory and their implementation in wireless systems. This textbook takes a unified view of the fundamentals of wireless communication and explains the web of concepts underpinning these advances at a level accessible to an audience with a basic background in probability and digital communication. Topics covered include MIMO (multiple input multiple output) communication, space-time coding, opportunistic communication, OFDM and CDMA. The concepts are illustrated using many examples from wireless systems such as GSM, IS-95 (CDMA), IS-856 (1× EV-DO), Flash OFDM and ArrayComm SDMA systems. Particular emphasis is placed on the interplay between concepts and their implementation in systems. An abundant supply of exercises and figures reinforce the material in the text. This book is intended for use on graduate courses in electrical and computer engineering and will also be of great interest to practicing engineers.

David Tse is a professor at the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley.

Pramod Viswanath is an assistant professor at the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign.

Fundamentals of Wireless Communication

David Tse
University of California, Berkeley

and

Pramod Viswanath
University of Illinois, Urbana-Champaign

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521845274

First published in print format 2005

ISBN-13 978-0-511-12598-0 eBook (NetLibrary)
ISBN-10 0-511-12598-4 eBook (NetLibrary)

ISBN-13 978-0-521-84527-4 hardback
ISBN-10 0-521-84527-0 hardback

To my family and Lavinia
DT

To my parents and to Suma
PV

Contents

Preface

Why we wrote this book

The writing of this book was prompted by two main developments in wireless communication in the past decade. First is the huge surge of research activities in physical-layer wireless communication theory. While this has been a subject of study since the sixties, recent developments such as opportunistic and multiple input multiple output (MIMO) communication techniques have brought completely new perspectives on how to communicate over wireless channels. Second is the rapid evolution of wireless systems, particularly cellular networks, which embody communication concepts of increasing sophistication. This evolution started with second-generation digital standards, particularly the IS-95 Code Division Multiple Access standard, continuing to more recent third-generation systems focusing on data applications. This book aims to present modern wireless communication concepts in a coherent and unified manner and to illustrate the concepts in the broader context of the wireless systems on which they have been applied.

Structure of the book

This book is a web of interlocking concepts. The concepts can be structured roughly into three levels:

1. channel characteristics and modeling;
2. communication concepts and techniques;
3. application of these concepts in a system context.

A wireless communication engineer should have an understanding of the concepts at all three levels as well as the tight interplay between the levels. We emphasize this interplay in the book by interlacing the chapters across these levels rather than presenting the topics sequentially from one level to the next.

- Chapter 2: basic properties of multipath wireless channels and their modeling (level 1).
- Chapter 3: point-to-point communication techniques that increase reliability by exploiting time, frequency and spatial diversity (2).
- Chapter 4: cellular system design via a case study of three systems, focusing on multiple access and interference management issues (3).
- Chapter 5: point-to-point communication revisited from a more fundamental capacity point of view, culminating in the modern concept of opportunistic communication (2).
- Chapter 6: multiuser capacity and opportunistic communication, and its application in a third-generation wireless data system (3).
- Chapter 7: MIMO channel modeling (1).
- Chapter 8: MIMO capacity and architectures (2).
- Chapter 9: diversity–multiplexing tradeoff and space-time code design (2).
- Chapter 10: MIMO in multiuser channels and cellular systems (3).

How to use this book

This book is written as a textbook for a first-year graduate course in wireless communication. The expected background is solid undergraduate/beginning graduate courses in signals and systems, probability and digital communication. This background is supplemented by the two appendices in the book. Appendix A summarizes some basic facts in vector detection and estimation in Gaussian noise which are used repeatedly throughout the book. Appendix B covers the underlying information theory behind the channel capacity results used in this book. Even though information theory has played a significant role in many of the recent developments in wireless communication, in the main text we only introduce capacity results in a heuristic manner and use them mainly to motivate communication concepts and techniques. No background in information theory is assumed. The appendix is intended for the reader who wants to have a more in-depth and unified understanding of the capacity results.

At Berkeley and Urbana-Champaign, we have used earlier versions of this book to teach one-semester (15 weeks) wireless communication courses. We have been able to cover most of the materials in Chapters 1 through 8 and parts of 9 and 10. Depending on the background of the students and the time available, one can envision several other ways to structure a course around this book. Examples:

- A senior level advanced undergraduate course in wireless communication: Chapters 2, 3, 4.
- An advanced graduate course for students with background in wireless channels and systems: Chapters 3, 5, 6, 7, 8, 9, 10.

- A short (quarter) course focusing on MIMO and space-time coding: Chapters 3, 5, 7, 8, 9.

The more than 230 exercises form an integral part of the book. Working on at least some of them is essential in understanding the material. Most of them elaborate on concepts discussed in the main text. The exercises range from relatively straightforward derivations of results in the main text, to "back-of-envelope" calculations for actual wireless systems, to "get-your-hands-dirty" MATLAB types, and to reading exercises that point to current research literature. The small bibliographical notes at the end of each chapter provide pointers to literature that is very closely related to the material discussed in the book; we do not aim to exhaust the immense research literature related to the material covered here.

Acknowledgements

We would like first to thank the students in our research groups for the selfless help they provided. In particular, many thanks to: Sanket Dusad, Raúl Etkin and Lenny Grokop, who between them painstakingly produced most of the figures in the book; Aleksandar Jovičić, who drew quite a few figures and proofread some chapters; Ada Poon whose research shaped significantly the material in Chapter 7 and who drew several figures in that chapter as well as in Chapter 2; Saurabha Tavildar and Lizhong Zheng whose research led to Chapter 9; Tie Liu and Vinod Prabhakaran for their help in clarifying and improving the presentation of Costa precoding in Chapter 10.

Several researchers read drafts of the book carefully and provided us with very useful comments on various chapters of the book: thanks to Stark Draper, Atilla Eryilmaz, Irem Koprulu, Dana Porrat and Pascal Vontobel. This book has also benefited immensely from critical comments from students who have taken our wireless communication courses at Berkeley and Urbana-Champaign. In particular, sincere thanks to Amir Salman Avestimehr, Alex Dimakis, Krishnan Eswaran, Jana van Greunen, Nils Hoven, Shridhar Mubaraq Mishra, Jonathan Tsao, Aaron Wagner, Hua Wang, Xinzhou Wu and Xue Yang.

Earlier drafts of this book have been used in teaching courses at several universities: Cornell, ETHZ, MIT, Northwestern and University of Colorado at Boulder. We would like to thank the instructors for their feedback: Helmut Bölcskei, Anna Scaglione, Mahesh Varanasi, Gregory Wornell and Lizhong Zheng. We would like to thank Ateet Kapur, Christian Peel and Ulrich Schuster from Helmut's group for their very useful feedback. Thanks are also due to Mitchell Trott for explaining to us how the ArrayComm systems work.

This book contains the results of many researchers, but it owes an intellectual debt to two individuals in particular. Bob Gallager's research and teaching style have greatly inspired our writing of this book. He has taught us that good theory, by providing a unified and conceptually simple understanding of a morass of results, should *shrink* rather than *grow* the knowledge tree. This book is an attempt to implement this dictum. Our many discussions with

Rajiv Laroia have significantly influenced our view of the system aspects of wireless communication. Several of his ideas have found their way into the “system view” discussions in the book.

Finally we would like to thank the National Science Foundation, whose continual support of our research led to this book.

Notation

Some specific sets

$\mathcal{R}$	Real numbers
$\mathcal{C}$	Complex numbers
$\mathcal{S}$	A subset of the users in the uplink of a cell

Scalars

m	Non-negative integer representing discrete-time
L	Number of diversity branches
ℓ	Scalar, indexing the diversity branches
K	Number of users
N	Block length
N_{c}	Number of tones in an OFDM system
T_{c}	Coherence time
T_{d}	Delay spread
W	Bandwidth
n_{t}	Number of transmit antennas
n_{r}	Number of receive antennas
n_{min}	Minimum of number of transmit and receive antennas
$h[m]$	Scalar channel, complex valued, at time m
h^*	Complex conjugate of the complex valued scalar h
$x[m]$	Channel input, complex valued, at time m
$y[m]$	Channel output, complex valued, at time m
$\mathcal{N}(\mu, \sigma^2)$	Real Gaussian random variable with mean μ and variance σ^2
$\mathcal{CN}(0, \sigma^2)$	Circularly symmetric complex Gaussian random variable: the real and imaginary parts are i.i.d. $\mathcal{N}(0, \sigma^2/2)$
N_0	Power spectral density of white Gaussian noise
$\{w[m]\}$	Additive Gaussian noise process, i.i.d. $\mathcal{CN}(0, N_0)$ with time m
$z[m]$	Additive colored Gaussian noise, at time m
P	Average power constraint measured in joules/symbol
$\bar{P}$	Average power constraint measured in watts
SNR	Signal-to-noise ratio
SINR	Signal-to-interference-plus-noise ratio

$\mathcal{E}_{\text{b}}$	Energy per received bit
P_{e}	Error probability

Capacities

C_{awgn}	Capacity of the additive white Gaussian noise channel
C_{ϵ}	ϵ-Outage capacity of the slow fading channel
C_{sum}	Sum capacity of the uplink or the downlink
C_{sym}	Symmetric capacity of the uplink or the downlink
$C_{\epsilon}^{\text{sym}}$	ϵ-Outage symmetric capacity of the slow fading uplink channel
p_{out}	Outage probability of a scalar fading channel
$p_{\text{out}}^{\text{Ala}}$	Outage probability when employing the Alamouti scheme
$p_{\text{out}}^{\text{rep}}$	Outage probability with the repetition scheme
$p_{\text{out}}^{\text{ul}}$	Outage probability of the uplink
$p_{\text{out}}^{\text{mimo}}$	Outage probability of the MIMO fading channel
$p_{\text{out}}^{\text{ul—mimo}}$	Outage probability of the uplink with multiple antennas at the base-station

Vectors and matrices

$\mathbf{h}$	Vector, complex valued, channel
$\mathbf{x}$	Vector channel input
$\mathbf{y}$	Vector channel output
$\mathcal{CN}(0, \mathbf{K})$	Circularly symmetric Gaussian random vector with mean zero and covariance matrix $\mathbf{K}$
$\mathbf{w}$	Additive Gaussian noise vector $\mathcal{CN}(0, N_0\mathbf{I})$
$\mathbf{h}^*$	Complex conjugate-transpose of $\mathbf{h}$
$\mathbf{d}$	Data vector
$\tilde{\mathbf{d}}$	Discrete Fourier transform of $\mathbf{d}$
$\mathbf{H}$	Matrix, complex valued, channel
$\mathbf{K}_x$	Covariance matrix of the random complex vector $\mathbf{x}$
$\mathbf{H}^*$	Complex conjugate-transpose of $\mathbf{H}$
$\mathbf{H}^t$	Transpose of matrix $\mathbf{H}$
$\mathbf{Q}, \mathbf{U}, \mathbf{V}$	Unitary matrices
$\mathbf{I}_n$	Identity $n \times n$ matrix
Λ, Ψ	Diagonal matrices
$\text{diag}\{p_1, \ldots, p_n\}$	Diagonal matrix with the diagonal entries equal to $p_1, \ldots, p_n$
$\mathbf{C}$	Circulant matrix
$\mathbf{D}$	Normalized codeword difference matrix

Operations

$\mathbb{E}[x]$	Mean of the random variable x
$\mathbb{P}\{A\}$	Probability of an event A
$\text{Tr}[\mathbf{K}]$	Trace of the square matrix $\mathbf{K}$
$\text{sinc}(t)$	Defined to be the ratio of $\sin(\pi t)$ to πt
$Q(a)$	$\int_a^\infty (1/\sqrt{2\pi}) \exp^{-x^2/2}\,\mathrm{d}x$
$\mathcal{L}(\cdot, \cdot)$	Lagrangian function

CHAPTER

1 Introduction

1.1 Book objective

Wireless communication is one of the most vibrant areas in the communication field today. While it has been a topic of study since the 1960s, the past decade has seen a surge of research activities in the area. This is due to a confluence of several factors. First, there has been an explosive increase in demand for tetherless connectivity, driven so far mainly by cellular telephony but expected to be soon eclipsed by wireless data applications. Second, the dramatic progress in VLSI technology has enabled small-area and low-power implementation of sophisticated signal processing algorithms and coding techniques. Third, the success of second-generation (2G) digital wireless standards, in particular, the IS-95 Code Division Multiple Access (CDMA) standard, provides a concrete demonstration that good ideas from communication theory can have a significant impact in practice. The research thrust in the past decade has led to a much richer set of perspectives and tools on how to communicate over wireless channels, and the picture is still very much evolving.

There are two fundamental aspects of wireless communication that make the problem challenging and interesting. These aspects are by and large not as significant in wireline communication. First is the phenomenon of *fading*: the time variation of the channel strengths due to the small-scale effect of multipath fading, as well as larger-scale effects such as path loss via distance attenuation and shadowing by obstacles. Second, unlike in the wired world where each transmitter–receiver pair can often be thought of as an isolated point-to-point link, wireless users communicate over the air and there is significant *interference* between them. The interference can be between transmitters communicating with a common receiver (e.g., uplink of a cellular system), between signals from a single transmitter to multiple receivers (e.g., downlink of a cellular system), or between different transmitter–receiver pairs (e.g., interference between users in different cells). How to deal with fading and with interference is central to the design of wireless communication systems and will be the central theme of this book. Although this book takes a physical-layer perspective, it will be seen that in fact the management of fading and interference has ramifications across multiple layers.

Traditionally the design of wireless systems has focused on increasing the *reliability* of the air interface; in this context, fading and interference are viewed as *nuisances* that are to be countered. Recent focus has shifted more towards increasing the *spectral efficiency*; associated with this shift is a new point of view that fading can be viewed as an *opportunity* to be exploited. The main objective of the book is to provide a unified treatment of wireless communication from both these points of view. In addition to traditional topics such as diversity and interference averaging, a substantial portion of the book will be devoted to more modern topics such as opportunistic and multiple input multiple output (MIMO) communication.

An important component of this book is the *system view* emphasis: the successful implementation of a theoretical concept or a technique requires an understanding of how it interacts with the wireless system as a whole. Unlike the derivation of a concept or a technique, this system view is less malleable to mathematical formulations and is primarily acquired through experience with designing actual wireless systems. We try to help the reader develop some of this intuition by giving numerous examples of how the concepts are applied in actual wireless systems. Five examples of wireless systems are used. The next section gives some sense of the scope of the wireless systems considered in this book.

1.2 Wireless systems

Wireless communication, despite the hype of the popular press, is a field that has been around for over a hundred years, starting around 1897 with Marconi's successful demonstrations of wireless telegraphy. By 1901, radio reception across the Atlantic Ocean had been established; thus, rapid progress in technology has also been around for quite a while. In the intervening hundred years, many types of wireless systems have flourished, and often later disappeared. For example, television transmission, in its early days, was broadcast by wireless radio transmitters, which are increasingly being replaced by cable transmission. Similarly, the point-to-point microwave circuits that formed the backbone of the telephone network are being replaced by optical fiber. In the first example, wireless technology became outdated when a wired distribution network was installed; in the second, a new wired technology (optical fiber) replaced the older technology. The opposite type of example is occurring today in telephony, where wireless (cellular) technology is partially replacing the use of the wired telephone network (particularly in parts of the world where the wired network is not well developed). The point of these examples is that there are many situations in which there is a choice between wireless and wire technologies, and the choice often changes when new technologies become available.

In this book, we will concentrate on cellular networks, both because they are of great current interest and also because the features of many other wireless systems can be easily understood as special cases or simple generalizations of the features of cellular networks. A cellular network consists of a large number of wireless subscribers who have cellular telephones (users), that can be used in cars, in buildings, on the street, or almost anywhere. There are also a number of fixed base-stations, arranged to provide coverage of the subscribers.

The area covered by a base-station, i.e., the area from which incoming calls reach that base-station, is called a cell. One often pictures a cell as a hexagonal region with the base-station in the middle. One then pictures a city or region as being broken up into a hexagonal lattice of cells (see Figure 1.1a). In reality, the base-stations are placed somewhat irregularly, depending on the location of places such as building tops or hill tops that have good communication coverage and that can be leased or bought (see Figure 1.1b). Similarly, mobile users connected to a base-station are chosen by good communication paths rather than geographic distance.

When a user makes a call, it is connected to the base-station to which it appears to have the best path (often but not always the closest base-station). The base-stations in a given area are then connected to a *mobile telephone switching office* (MTSO, also called a *mobile switching center* MSC) by high-speed wire connections or microwave links. The MTSO is connected to the public wired telephone network. Thus an incoming call from a mobile user is first connected to a base-station and from there to the MTSO and then to the wired network. From there the call goes to its destination, which might be an ordinary wire line telephone, or might be another mobile subscriber. Thus, we see that a cellular network is not an independent network, but rather an appendage to the wired network. The MTSO also plays a major role in coordinating which base-station will handle a call to or from a user and when to handoff a user from one base-station to another.

When another user (either wired or wireless) places a call to a given user, the reverse process takes place. First the MTSO for the called subscriber is found,

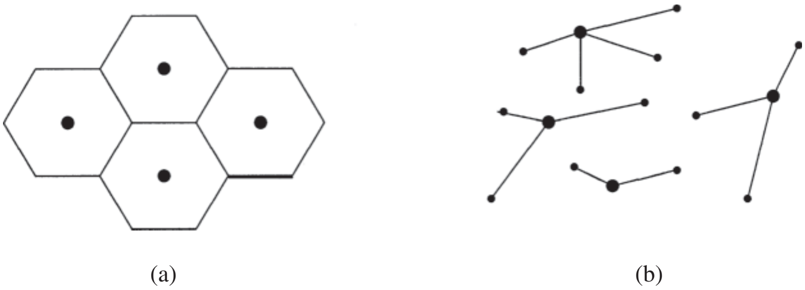


Figure 1.1 Cells and base-stations for a cellular network. (a) An oversimplified view in which each cell is hexagonal. (b) A more realistic case where base-stations are irregularly placed and cell phones choose the best base-station.

then the closest base-station is found, and finally the call is set up through the MTSO and the base-station. The wireless link from a base-station to the mobile users is interchangeably called the *downlink* or the *forward channel*, and the link from the users to a base-station is called the *uplink* or a *reverse channel*. There are usually many users connected to a single base-station, and thus, for the downlink channel, the base-station must multiplex together the signals to the various connected users and then broadcast one waveform from which each user can extract its own signal. For the uplink channel, each user connected to a given base-station transmits its own waveform, and the base-station receives the sum of the waveforms from the various users plus noise. The base-station must then separate out the signals from each user and forward these signals to the MTSO.

Older cellular systems, such as the AMPS (advanced mobile phone service) system developed in the USA in the eighties, are analog. That is, a voice waveform is modulated on a carrier and transmitted without being transformed into a digital stream. Different users in the same cell are assigned different modulation frequencies, and adjacent cells use different sets of frequencies. Cells sufficiently far away from each other can reuse the same set of frequencies with little danger of interference.

Second-generation cellular systems are digital. One is the GSM (global system for mobile communication) system, which was standardized in Europe but now used worldwide, another is the TDMA (time-division multiple access) standard developed in the USA (IS-136), and a third is CDMA (code division multiple access) (IS-95). Since these cellular systems, and their standards, were originally developed for telephony, the current data rates and delays in cellular systems are essentially determined by voice requirements. Third-generation cellular systems are designed to handle data and/or voice. While some of the third-generation systems are essentially evolution of second-generation voice systems, others are designed from scratch to cater for the specific characteristics of data. In addition to a requirement for higher rates, data applications have two features that distinguish them from voice:

- Many data applications are extremely bursty; users may remain inactive for long periods of time but have very high demands for short periods of time. Voice applications, in contrast, have a fixed-rate demand over long periods of time.
- Voice has a relatively tight latency requirement of the order of 100 ms. Data applications have a wide range of latency requirements; real-time applications, such as gaming, may have even tighter delay requirements than voice, while many others, such as http file transfers, have a much laxer requirement.

In the book we will see the impact of these features on the appropriate choice of communication techniques.

As mentioned above, there are many kinds of wireless systems other than cellular. First there are the broadcast systems such as AM radio, FM radio, TV and paging systems. All of these are similar to the downlink part of cellular networks, although the data rates, the sizes of the areas covered by each broadcasting node and the frequency ranges are very different. Next, there are wireless LANs (local area networks). These are designed for much higher data rates than cellular systems, but otherwise are similar to a single cell of a cellular system. These are designed to connect laptops and other portable devices in the local area network within an office building or similar environment. There is little mobility expected in such systems and their major function is to allow portability. The major standards for wireless LANs are the IEEE 802.11 family. There are smaller-scale standards like Bluetooth or a more recent one based on ultra-wideband (UWB) communication whose purpose is to reduce cabling in an office and simplify transfers between office and hand-held devices. Finally, there is another type of LAN called an *ad hoc network*. Here, instead of a central node (base-station) through which all traffic flows, the nodes are all alike. The network organizes itself into links between various pairs of nodes and develops routing tables using these links. Here the network layer issues of routing, dissemination of control information, etc. are important concerns, although problems of relaying and distributed cooperation between nodes can be tackled from the physical-layer as well and are active areas of current research.

1.3 Book outline

The central object of interest is the wireless fading channel. Chapter 2 introduces the multipath fading channel model that we use for the rest of the book. Starting from a continuous-time passband channel, we derive a discrete-time complex baseband model more suitable for analysis and design. Key physical parameters such as coherence time, coherence bandwidth, Doppler spread and delay spread are explained and several statistical models for multipath fading are surveyed. There have been many statistical models proposed in the literature; we will be far from exhaustive here. The goal is to have a small set of example models in our repertoire to evaluate the performance of basic communication techniques we will study.

Chapter 3 introduces many of the issues of communicating over fading channels in the simplest point-to-point context. As a baseline, we start by looking at the problem of detection of uncoded transmission over a narrowband fading channel. We find that the performance is very poor, much worse than over the additive white Gaussian noise (AWGN) channel with the same average signal-to-noise ratio (SNR). This is due to a significant probability that the channel is in *deep fade*. Various *diversity techniques* to mitigate this adverse effect of fading are then studied. Diversity techniques increase reliability by sending the same information through multiple independently faded paths so that the probability of successful transmission is higher. Some of the techniques studied include:

- interleaving of coded symbols over time to obtain time diversity;
- inter-symbol equalization, multipath combining in spread-spectrum systems and coding over sub-carriers in orthogonal frequency division multiplexing (OFDM) systems to obtain frequency diversity;
- use of multiple transmit and/or receive antennas, via *space-time* coding, to obtain spatial diversity.

In some scenarios, there is an interesting interplay between channel uncertainty and the diversity gain: as the number of diversity branches increases, the performance of the system first improves due to the diversity gain but then subsequently deteriorates as channel uncertainty makes it more difficult to combine signals from the different branches.

In Chapter 4 the focus is shifted from point-to-point communication to studying cellular systems as a whole. Multiple access and inter-cell interference management are the key issues that come to the forefront. We explain how existing digital wireless systems deal with these issues. The concepts of frequency reuse and cell sectorization are discussed, and we contrast narrowband systems such as GSM and IS-136, where users within the same cell are kept orthogonal and frequency is reused only in cells far away, and CDMA systems, such as IS-95, where the signals of users both within the same cell and across different cells are spread across the same spectrum, i.e., frequency reuse factor of 1. Due to the full reuse, CDMA systems have to manage intra-cell and inter-cell interference more efficiently: in addition to the diversity techniques of time-interleaving, multipath combining and soft handoff, *power control* and *interference averaging* are the key interference management mechanisms. All the five techniques strive toward the same system goal: to maintain the channel quality of each user, as measured by the signal-to-interference-and-noise ratio (SINR), as constant as possible. This chapter is concluded with the discussion of a wideband OFDM system, which combines the advantages of both the CDMA and the narrowband systems.

Chapter 5 studies the capacity of wireless channels. This provides a higher level view of the tradeoffs involved in the earlier chapters and also lays the foundation for understanding the more modern developments in the subsequent chapters. The performance over the (non-faded) AWGN channel, as a baseline for comparison. We introduce the concept of *channel capacity* as the basic performance measure. The capacity of a channel provides the fundamental limit of communication achievable by any scheme. For the fading channel, there are several capacity measures, relevant for different scenarios. Two distinct scenarios provide particular insight: (1) the *slow* fading channel, where the channel stays the same (random value) over the entire time-scale of communication, and (2) the *fast* fading channel, where the channel varies significantly over the time-scale of communication.

In the slow fading channel, the key event of interest is *outage*: this is the situation when the channel is so poor that no scheme can communicate reliably at a certain target data rate. The largest rate of reliable communication at a certain outage probability is called the outage capacity. In the fast fading channel, in contrast, outage can be avoided due to the ability to average over the time variation of the channel, and one can define a positive capacity at which arbitrarily reliable communication is possible. Using these capacity measures, several resources associated with a fading channel are defined: (1) diversity; (2) number of degrees of freedom; (3) received power. These three resources form a basis for assessing the nature of performance gain by the various communication schemes studied in the rest of the book.

Chapters 6 to 10 cover the more recent developments in the field. In Chapter 6 we revisit the problem of multiple access over fading channels from a more fundamental point of view. Information theory suggests that if both the transmitters and the receiver can track the fading channel, the optimal strategy to maximize the total system throughput is to allow only the user with the best channel to transmit at any time. A similar strategy is also optimal for the downlink. Opportunistic strategies of this type yield a system-wide *multiuser diversity* gain: the more users in the system, the larger the gain, as there is more likely to be a user with a very strong channel. To implement this concept in a real system, three important considerations are: *fairness* of the resource allocation across users; *delay* experienced by the individual user waiting for its channel to become good; and *measurement inaccuracy* and *delay* in feeding back the channel state to the transmitters. We discuss how these issues are addressed in the context of IS-865 (also called HDR or CDMA 2000 1× EV-DO), a third-generation wireless data system.

A wireless system consists of multiple dimensions: time, frequency, space and users. Opportunistic communication maximizes the spectral efficiency by measuring when and where the channel is good and only transmits in those degrees of freedom. In this context, channel fading is *beneficial* in the sense that the fluctuation of the channel across the degrees of freedom ensures that there will be some degrees of freedom in which the channel is very good. This is in sharp contrast to the diversity-based approach in Chapter 3, where channel fluctuation is always detrimental and the design goal is to average out the fading to make the overall channel as constant as possible. Taking this philosophy one step further, we discuss a technique, called *opportunistic beamforming*, in which channel fluctuation can be *induced* in situations when the natural fading has small dynamic range and/or is slow. From the cellular system point of view, this technique also increases the fluctuations of the *interference* imparted on adjacent cells, and presents an opposing philosophy to the notion of interference averaging in CDMA systems.

Chapters 7, 8, 9 and 10 discuss multiple input multiple output (MIMO) communication. It has been known for a while that the uplink with multiple receive antennas at the base-station allow several users to simultaneously communicate to the receiver. The multiple antennas in effect increase the number of degrees of freedom in the system and allow spatial separation of the signals from the different users. It has recently been shown that a similar effect occurs for point-to-point channels with multiple transmit *and* receive antennas, i.e., even when the antennas of the multiple users are co-located. This holds provided that the scattering environment is rich enough to allow the receive antennas to separate out the signal from the different transmit antennas, allowing the *spatial multiplexing* of information. This is yet another example where channel fading is beneficial to communication. Chapter 7 studies the properties of the multipath environment that determine the amount of spatial multiplexing possible and defines an *angular domain* in which such properties are seen most explicitly. We conclude with a class of statistical MIMO channel models, based in the angular domain, which will be used in later chapters to analyze the performance of communication techniques.

Chapter 8 discusses the capacity and capacity-achieving transceiver architectures for MIMO channels, focusing on the fast fading scenario. It is demonstrated that the fast fading capacity increases linearly with the minimum of the number of transmit and receive antennas at all values of SNR. At high SNR, the linear increase is due to the increase in degrees of freedom from spatial multiplexing. At low SNR, the linear increase is due to a power gain from receive beamforming. At intermediate SNR ranges, the linear increase is due to a combination of both these gains. Next, we study the transceiver architectures that achieve the capacity of the fast fading channel. The focus is on the V-BLAST architecture, which multiplexes independent data streams, one onto each of the transmit antennas. A variety of receiver structures are considered: these include the decorrelator and the linear minimum mean square-error (MMSE) receiver. The performance of these receivers can be enhanced by successively canceling the streams as they are decoded; this is known as successive interference cancellation (SIC). It is shown that the MMSE–SIC receiver achieves the capacity of the fast fading MIMO channel.

The V-BLAST architecture is very suboptimal for the slow fading MIMO channel: it does not code across the transmit antennas and thus the diversity gain is limited by that obtained with the receive antenna array. A modification, called D-BLAST, where the data streams are *interleaved* across the transmit antenna array, achieves the outage capacity of the slow fading MIMO channel. The boost of the outage capacity of a MIMO channel as compared to a single antenna channel is due to a combination of both diversity and spatial multiplexing gains. In Chapter 9, we study a fundamental *tradeoff* between the diversity and multiplexing gains that can be simultaneously harnessed over a slow fading MIMO channel. This formulation is then used as a unified framework to assess both the diversity and multiplexing performance of several schemes that have appeared earlier in the book. This framework is also used to motivate the construction of new tradeoff-optimal space-time codes. In particular, we discuss an approach to design *universal* space-time codes that are tradeoff-optimal.

Finally, Chapter 10 studies the use of multiple transmit and receive antennas in multiuser and cellular systems; this is also called *space-division multiple access* (SDMA). Here, in addition to providing spatial multiplexing and diversity, multiple antennas can also be used to mitigate interference between different users. In the uplink, interference mitigation is done at the base-station via the SIC receiver. In the downlink, interference mitigation is also done at the base-station and this requires *precoding*: we study a precoding scheme, called Costa or dirty-paper precoding, that is the natural analog of the SIC receiver in the uplink. This study allows us to relate the performance of an SIC receiver in the uplink with a corresponding precoding scheme in a *reciprocal* downlink. The ArrayComm system is used as an example of an SDMA cellular system.

CHAPTER

2 The wireless channel

A good understanding of the wireless channel, its key physical parameters and the modeling issues, lays the foundation for the rest of the book. This is the goal of this chapter.

A defining characteristic of the mobile wireless channel is the variations of the channel strength over time and over frequency. The variations can be roughly divided into two types (Figure 2.1):

- *Large-scale fading*, due to path loss of signal as a function of distance and shadowing by large objects such as buildings and hills. This occurs as the mobile moves through a distance of the order of the cell size, and is typically frequency independent.
- *Small-scale fading*, due to the constructive and destructive interference of the multiple signal paths between the transmitter and receiver. This occurs at the spatial scale of the order of the carrier wavelength, and is frequency dependent.

We will talk about both types of fading in this chapter, but with more emphasis on the latter. Large-scale fading is more relevant to issues such as cell-site planning. Small-scale multipath fading is more relevant to the design of reliable and efficient communication systems – the focus of this book.

We start with the physical modeling of the wireless channel in terms of electromagnetic waves. We then derive an input/output linear time-varying model for the channel, and define some important physical parameters. Finally, we introduce a few statistical models of the channel variation over time and over frequency.

2.1 Physical modeling for wireless channels

Wireless channels operate through electromagnetic radiation from the transmitter to the receiver. In principle, one could solve the electromagnetic field equations, in conjunction with the transmitted signal, to find the

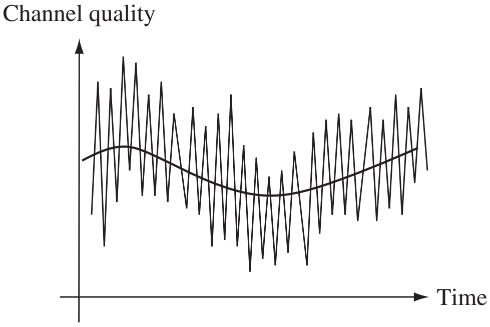


Figure 2.1 Channel quality varies over multiple time-scales. At a slow scale, channel varies due to large-scale fading effects. At a fast scale, channel varies due to multipath effects.

electromagnetic field impinging on the receiver antenna. This would have to be done taking into account the obstructions caused by ground, buildings, vehicles, etc. in the vicinity of this electromagnetic wave.[1]

Cellular communication in the USA is limited by the Federal Communication Commission (FCC), and by similar authorities in other countries, to one of three frequency bands, one around 0.9 GHz, one around 1.9 GHz, and one around 5.8 GHz. The wavelength λ of electromagnetic radiation at any given frequency f is given by $\lambda = c/f$, where $c = 3 \times 10^8$ m/s is the speed of light. The wavelength in these cellular bands is thus a fraction of a meter, so to calculate the electromagnetic field at a receiver, the locations of the receiver and the obstructions would have to be known within sub-meter accuracies. The electromagnetic field equations are therefore too complex to solve, especially on the fly for mobile users. Thus, we have to ask what we really need to know about these channels, and what approximations might be reasonable.

One of the important questions is where to choose to place the base-stations, and what range of power levels are then necessary on the downlink and uplink channels. To some extent this question must be answered experimentally, but it certainly helps to have a sense of what types of phenomena to expect. Another major question is what types of modulation and detection techniques look promising. Here again, we need a sense of what types of phenomena to expect. To address this, we will construct stochastic models of the channel, assuming that different channel behaviors appear with different probabilities, and change over time (with specific stochastic properties). We will return to the question of why such stochastic models are appropriate, but for now we simply want to explore the gross characteristics of these channels. Let us start by looking at several over-idealized examples.

[1] By obstructions, we mean not only objects in the line-of-sight between transmitter and receiver, but also objects in locations that cause non-negligible changes in the electromagnetic field at the receiver; we shall see examples of such obstructions later.

2.1.1 Free space, fixed transmit and receive antennas

First consider a fixed antenna radiating into free space. In the far field,[2] the electric field and magnetic field at any given location are perpendicular both to each other and to the direction of propagation from the antenna. They are also proportional to each other, so it is sufficient to know only one of them (just as in wired communication, where we view a signal as simply a voltage waveform or a current waveform). In response to a transmitted sinusoid $\cos 2\pi ft$, we can express the electric far field at time t as

$$E(f, t, (r, \theta, \psi)) = \frac{\alpha_s(\theta, \psi, f) \cos 2\pi f(t - r/c)}{r}. \tag{2.1}$$

Here, (r, θ, ψ) represents the point $\mathbf{u}$ in space at which the electric field is being measured, where r is the distance from the transmit antenna to $\mathbf{u}$ and where (θ, ψ) represents the vertical and horizontal angles from the antenna to $\mathbf{u}$ respectively. The constant c is the speed of light, and $\alpha_s(\theta, \psi, f)$ is the radiation pattern of the sending antenna at frequency f in the direction (θ, ψ); it also contains a scaling factor to account for antenna losses. Note that the phase of the field varies with fr/c, corresponding to the delay caused by the radiation traveling at the speed of light.

We are not concerned here with actually finding the radiation pattern for any given antenna, but only with recognizing that antennas have radiation patterns, and that the free space far field behaves as above.

It is important to observe that, as the distance r increases, the electric field decreases as r^{-1} and thus the power per square meter in the free space wave decreases as r^{-2}. This is expected, since if we look at concentric spheres of increasing radius r around the antenna, the total power radiated through the sphere remains constant, but the surface area increases as r^2. Thus, the power per unit area must decrease as r^{-2}. We will see shortly that this r^{-2} reduction of power with distance is often not valid when there are obstructions to free space propagation.

Next, suppose there is a fixed receive antenna at the location $\mathbf{u} = (r, \theta, \psi)$. The received waveform (in the absence of noise) in response to the above transmitted sinusoid is then

$$E_r(f, t, \mathbf{u}) = \frac{\alpha(\theta, \psi, f) \cos 2\pi f(t - r/c)}{r}, \tag{2.2}$$

where $\alpha(\theta, \psi, f)$ is the product of the antenna patterns of transmit and receive antennas in the given direction. Our approach to (2.2) is a bit odd since we started with the free space field at $\mathbf{u}$ in the absence of an antenna. Placing a

[2] The far field is the field sufficiently far away from the antenna so that (2.1) is valid. For cellular systems, it is a safe assumption that the receiver is in the far field.

receive antenna there changes the electric field in the vicinity of $\mathbf{u}$, but this is taken into account by the antenna pattern of the receive antenna.

Now suppose, for the given $\mathbf{u}$, that we define

$$H(f) := \frac{\alpha(\theta, \psi, f)\mathrm{e}^{-\mathrm{j}2\pi fr/c}}{r}. \tag{2.3}$$

We then have $E_{\mathrm{r}}(f, t, \mathbf{u}) = \Re\left[H(f)\mathrm{e}^{\mathrm{j}2\pi ft}\right]$. We have not mentioned it yet, but (2.1) and (2.2) are both linear in the input. That is, the received field (waveform) at $\mathbf{u}$ in response to a weighted sum of transmitted waveforms is simply the weighted sum of responses to those individual waveforms. Thus, $H(f)$ is the system function for an LTI (linear time-invariant) channel, and its inverse Fourier transform is the impulse response. The need for understanding electromagnetism is to determine what this system function is. We will find in what follows that linearity is a good assumption for all the wireless channels we consider, but that the time invariance does not hold when either the antennas or obstructions are in relative motion.

2.1.2 Free space, moving antenna

Next consider the fixed antenna and free space model above with a receive antenna that is moving with speed v in the direction of increasing distance from the transmit antenna. That is, we assume that the receive antenna is at a moving location described as $\mathbf{u}(t) = (r(t), \theta, \psi)$ with $r(t) = r_0 + vt$. Using (2.1) to describe the free space electric field at the moving point $\mathbf{u}(t)$ (for the moment with no receive antenna), we have

$$E(f, t, (r_0 + vt, \theta, \psi)) = \frac{\alpha_{\mathrm{s}}(\theta, \psi, f)\cos 2\pi f(t - r_0/c - vt/c)}{r_0 + vt}. \tag{2.4}$$

Note that we can rewrite $f(t - r_0/c - vt/c)$ as $f(1 - v/c)t - fr_0/c$. Thus, the sinusoid at frequency f has been converted to a sinusoid of frequency $f(1 - v/c)$; there has been a *Doppler shift* of $-fv/c$ due to the motion of the observation point.[3] Intuitively, each successive crest in the transmitted sinusoid has to travel a little further before it gets observed at the moving observation point. If the antenna is now placed at $\mathbf{u}(t)$, and the change of field due to the antenna presence is again represented by the receive antenna pattern, the received waveform, in analogy to (2.2), is

$$E_{\mathrm{r}}(f, t, (r_0 + vt, \theta, \psi)) = \frac{\alpha(\theta, \psi, f)\cos 2\pi f[(1 - v/c)t - r_0/c]}{r_0 + vt}. \tag{2.5}$$

[3] The reader should be familiar with the Doppler shift associated with moving cars. When an ambulance is rapidly moving toward us we hear a higher frequency siren. When it passes us we hear a rapid shift toward a lower frequency.

This channel cannot be represented as an LTI channel. If we ignore the time-varying attenuation in the denominator of (2.5), however, we can represent the channel in terms of a system function followed by translating the frequency f by the Doppler shift $-fv/c$. It is important to observe that the amount of shift depends on the frequency f. We will come back to discussing the importance of this Doppler shift and of the time-varying attenuation after considering the next example.

The above analysis does not depend on whether it is the transmitter or the receiver (or both) that are moving. So long as $r(t)$ is interpreted as the distance between the antennas (and the relative orientations of the antennas are constant), (2.4) and (2.5) are valid.

2.1.3 Reflecting wall, fixed antenna

Consider Figure 2.2 in which there is a fixed antenna transmitting the sinusoid $\cos 2\pi ft$, a fixed receive antenna, and a single perfectly reflecting large fixed wall. We assume that in the absence of the receive antenna, the electromagnetic field at the point where the receive antenna will be placed is the sum of the free space field coming from the transmit antenna plus a reflected wave coming from the wall. As before, in the presence of the receive antenna, the perturbation of the field due to the antenna is represented by the antenna pattern. An additional assumption here is that the presence of the receive antenna does not appreciably affect the plane wave impinging on the wall. In essence, what we have done here is to approximate the solution of Maxwell's equations by a method called *ray tracing*. The assumption here is that the received waveform can be approximated by the sum of the free space wave from the transmitter plus the reflected free space waves from each of the reflecting obstacles.

In the present situation, if we assume that the wall is very large, the reflected wave at a given point is the same (except for a sign change[4]) as the free space wave that would exist on the opposite side of the wall if the wall were not present (see Figure 2.3). This means that the reflected wave from the wall has the intensity of a free space wave at a distance equal to the distance to the wall and then

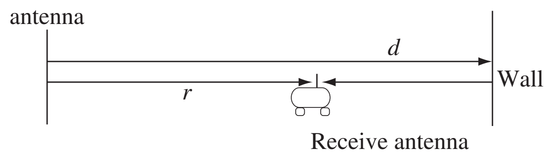


Figure 2.2 Illustration of a direct path and a reflected path.

[4] By basic electromagnetics, this sign change is a consequence of the fact that the electric field is parallel to the plane of the wall for this example.

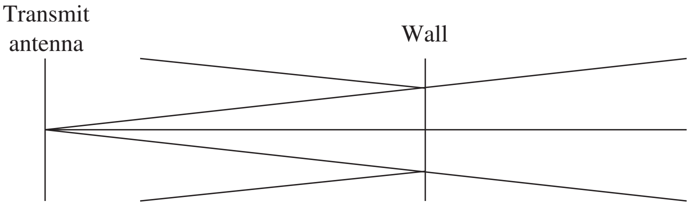


Figure 2.3 Relation of reflected wave to wave without wall.

back to the receive antenna, i.e., $2d-r$. Using (2.2) for both the direct and the reflected wave, and assuming the same antenna gain α for both waves, we get

$$E_{\mathrm{r}}(f,t)=\frac{\alpha\cos 2\pi f(t-r/c)}{r}-\frac{\alpha\cos 2\pi f(t-(2d-r)/c)}{2d-r}. \tag{2.6}$$

The received signal is a superposition of two waves, both of frequency f. The phase difference between the two waves is

$$\Delta\theta=\left(\frac{2\pi f(2d-r)}{c}+\pi\right)-\left(\frac{2\pi fr}{c}\right)=\frac{4\pi f}{c}(d-r)+\pi. \tag{2.7}$$

When the phase difference is an integer multiple of 2π, the two waves add *constructively*, and the received signal is strong. When the phase difference is an odd integer multiple of π, the two waves add *destructively*, and the received signal is weak. As a function of r, this translates into a spatial pattern of constructive and destructive interference of the waves. The distance from a peak to a valley is called the *coherence distance*:

$$\boxed{\Delta x_{\mathrm{c}}:=\frac{\lambda}{4},} \tag{2.8}$$

where $\lambda := c/f$ is the wavelength of the transmitted sinusoid. At distances much smaller than Δx_{c}, the received signal at a particular time does not change appreciably.

The constructive and destructive interference pattern also depends on the frequency f: for a fixed r, if f changes by

$$\frac{1}{2}\left(\frac{2d-r}{c}-\frac{r}{c}\right)^{-1}, \tag{2.9}$$

we move from a peak to a valley. The quantity

$$\boxed{T_{\mathrm{d}}:=\frac{2d-r}{c}-\frac{r}{c}} \tag{2.10}$$

is called the *delay spread* of the channel: it is the difference between the propagation delays along the two signal paths. The constructive and destructive interference pattern does not change appreciably if the frequency changes by an amount much smaller than $1/T_{\mathrm{d}}$. This parameter is called the *coherence bandwidth*.

2.1.4 Reflecting wall, moving antenna

Suppose the receive antenna is now moving at a velocity v (Figure 2.4). As it moves through the pattern of constructive and destructive interference created by the two waves, the strength of the received signal increases and decreases. This is the phenomenon of *multipath fading*. The time taken to travel from a peak to a valley is $c/(4fv)$: this is the time-scale at which the fading occurs, and it is called the *coherence time* of the channel.

An equivalent way of seeing this is in terms of the Doppler shifts of the direct and the reflected waves. Suppose the receive antenna is at location r_0 at time 0. Taking $r = r_0 + vt$ in (2.6), we get

$$\begin{aligned} E_r(f, t) = & \frac{\alpha \cos 2\pi f[(1 - v/c)t - r_0/c]}{r_0 + vt} \\ & - \frac{\alpha \cos 2\pi f[(1 + v/c)t + (r_0 - 2d)/c]}{2d - r_0 - vt}. \end{aligned} \tag{2.11}$$

The first term, the direct wave, is a sinusoid at frequency $f(1 - v/c)$, experiencing a Doppler shift $D_1 := -fv/c$. The second is a sinusoid at frequency $f(1 + v/c)$, with a Doppler shift $D_2 := +fv/c$. The parameter

$$\boxed{D_s := D_2 - D_1} \tag{2.12}$$

is called the *Doppler spread*. For example, if the mobile is moving at 60 km/h and $f = 900$ MHz, the Doppler spread is 100 Hz. The role of the Doppler spread can be visualized most easily when the mobile is much closer to the wall than to the transmit antenna. In this case the attenuations are roughly the same for both paths, and we can approximate the denominator of the second term by $r = r_0 + vt$. Then, combining the two sinusoids, we get

$$E_r(f, t) \approx \frac{2\alpha \sin 2\pi f\, [vt/c + (r_0 - d)/c] \sin 2\pi f[t - d/c]}{r_0 + vt}. \tag{2.13}$$

This is the product of two sinusoids, one at the input frequency f, which is typically of the order of GHz, and the other one at $fv/c = D_s/2$, which might be of the order of 50 Hz. Thus, the response to a sinusoid at f is another sinusoid at f with a time-varying envelope, with peaks going to zeros around every 5 ms (Figure 2.5). The envelope is at its widest when the mobile is at a peak of the

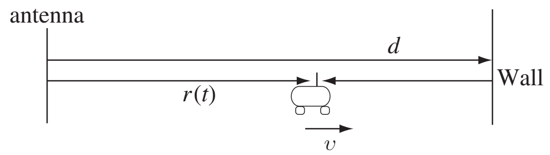


Figure 2.4 Illustration of a direct path and a reflected path.

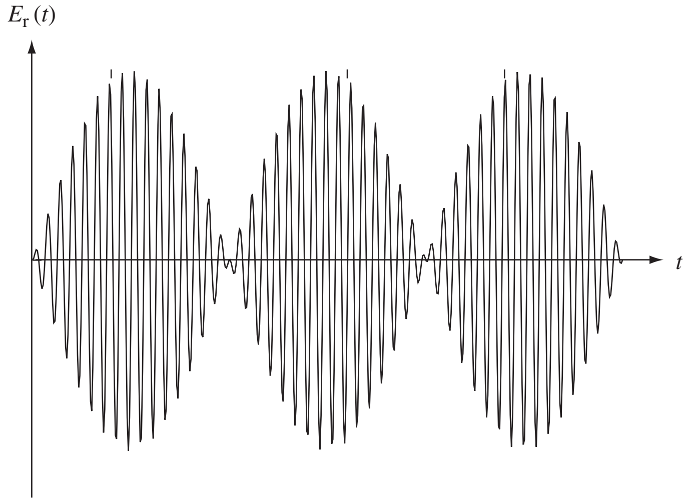


Figure 2.5 The received waveform oscillating at frequency f with a slowly varying envelope at frequency $D_s/2$.

interference pattern and at its narrowest when the mobile is at a valley. Thus, the Doppler spread determines the rate of traversal across the interference pattern and is inversely proportional to the coherence time of the channel.

We now see why we have partially ignored the denominator terms in (2.11) and (2.13). When the difference in the length between two paths changes by a quarter wavelength, the phase difference between the responses on the two paths changes by $\pi/2$, which causes a very significant change in the overall received amplitude. Since the carrier wavelength is very small relative to the path lengths, the time over which this phase effect causes a significant change is far smaller than the time over which the denominator terms cause a significant change. The effect of the phase changes is of the order of milliseconds, whereas the effect of changes in the denominator is of the order of seconds or minutes. In terms of modulation and detection, the time-scales of interest are in the range of milliseconds and less, and the denominators are effectively constant over these periods.

The reader might notice that we are constantly making approximations in trying to understand wireless communication, much more so than for wired communication. This is partly because wired channels are typically time-invariant over a very long time-scale, while wireless channels are typically time-varying, and appropriate models depend very much on the time-scales of interest. For wireless systems, the most important issue is what approximations to make. Thus, it is important to understand these modeling issues thoroughly.

2.1.5 Reflection from a ground plane

Consider a transmit and a receive antenna, both above a plane surface such as a road (Figure 2.6). When the horizontal distance r between the antennas becomes very large relative to their vertical displacements from the ground

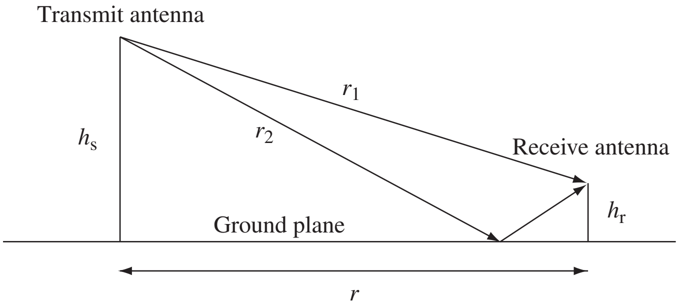


Figure 2.6 Illustration of a direct path and a reflected path off a ground plane.

plane (i.e., height), a very surprising thing happens. In particular, the difference between the direct path length and the reflected path length goes to zero as r^{-1} with increasing r (Exercise 2.5). When r is large enough, this difference between the path lengths becomes small relative to the wavelength c/f. Since the sign of the electric field is reversed on the reflected path[5], these two waves start to cancel each other out. The electric wave at the receiver is then attenuated as r^{-2}, and the received power decreases as r^{-4}. This situation is particularly important in rural areas where base-stations tend to be placed on roads.

2.1.6 Power decay with distance and shadowing

The previous example with reflection from a ground plane suggests that the received power can decrease with distance faster than r^{-2} in the presence of disturbances to free space. In practice, there are several obstacles between the transmitter and the receiver and, further, the obstacles might also absorb some power while scattering the rest. Thus, one expects the power decay to be considerably faster than r^{-2}. Indeed, empirical evidence from experimental field studies suggests that while power decay near the transmitter is like r^{-2}, at large distances the power can even decay *exponentially* with distance.

The ray tracing approach used so far provides a high degree of numerical accuracy in determining the electric field at the receiver, but requires a precise physical model including the location of the obstacles. But here, we are only looking for the order of decay of power with distance and can consider an alternative approach. So we look for a model of the physical environment with the fewest parameters but one that still provides useful global information about the field properties. A simple probabilistic model with two parameters of the physical environment, the density of the obstacles and the fraction of energy each object absorbs, is developed in Exercise 2.6. With each obstacle

[5] This is clearly true if the electric field is parallel to the ground plane. It turns out that this is also true for arbitrary orientations of the electric field, as long as the ground is not a perfect conductor and the angle of incidence is small enough. The underlying electromagnetics is analyzed in Chapter 2 of Jakes [62].

absorbing the same fraction of the energy impinging on it, the model allows us to show that the power decays exponentially in distance at a rate that is proportional to the density of the obstacles.

With a limit on the transmit power (either at the base-station or at the mobile), the largest distance between the base-station and a mobile at which communication can reliably take place is called the *coverage* of the cell. For reliable communication, a minimal received power level has to be met and thus the fast decay of power with distance constrains cell coverage. On the other hand, rapid signal attenuation with distance is also helpful; it reduces the *interference* between adjacent cells. As cellular systems become more popular, however, the major determinant of cell size is the number of mobiles in the cell. In engineering jargon, the cell is said to be *capacity* limited instead of coverage limited. The size of cells has been steadily decreasing, and one talks of micro cells and pico cells as a response to this effect. With capacity limited cells, the inter-cell interference may be intolerably high. To alleviate the inter-cell interference, neighboring cells use different parts of the frequency spectrum, and frequency is reused at cells that are far enough. Rapid signal attenuation with distance allows frequencies to be reused at closer distances.

The density of obstacles between the transmit and receive antennas depends very much on the physical environment. For example, outdoor plains have very little by way of obstacles while indoor environments pose many obstacles. This randomness in the environment is captured by modeling the density of obstacles and their absorption behavior as random numbers; the overall phenomenon is called *shadowing*.[6] The effect of shadow fading differs from multipath fading in an important way. The duration of a shadow fade lasts for multiple seconds or minutes, and hence occurs at a much slower time-scale compared to multipath fading.

2.1.7 Moving antenna, multiple reflectors

Dealing with multiple reflectors, using the technique of ray tracing, is in principle simply a matter of modeling the received waveform as the sum of the responses from the different paths rather than just two paths. We have seen enough examples, however, to understand that finding the magnitudes and phases of these responses is no simple task. Even for the very simple large wall example in Figure 2.2, the reflected field calculated in (2.6) is valid only at distances from the wall that are small relative to the dimensions of the wall. At very large distances, the total power reflected from the wall is proportional to both d^{-2} and to the area of the cross section of the wall. The power reaching the receiver is proportional to $(d - r(t))^{-2}$. Thus, the power attenuation from transmitter to receiver (for the large distance case) is proportional to $(d(d - r(t)))^{-2}$ rather

[6] This is called shadowing because it is similar to the effect of clouds partly blocking sunlight.

than to $(2d - r(t))^{-2}$. This shows that ray tracing must be used with some caution. Fortunately, however, linearity still holds in these more complex cases.

Another type of reflection is known as *scattering* and can occur in the atmosphere or in reflections from very rough objects. Here there are a very large number of individual paths, and the received waveform is better modeled as an integral over paths with infinitesimally small differences in their lengths, rather than as a sum.

Knowing how to find the amplitude of the reflected field from each type of reflector is helpful in determining the coverage of a base-station (although ultimately experimentation is necessary). This is an important topic if our objective is trying to determine where to place base-stations. Studying this in more depth, however, would take us afield and too far into electromagnetic theory. In addition, we are primarily interested in questions of modulation, detection, multiple access, and network protocols rather than location of base-stations. Thus, we turn our attention to understanding the nature of the aggregate received waveform, given a representation for each reflected wave. This leads to modeling the input/output behavior of a channel rather than the detailed response on each path.

2.2 Input/output model of the wireless channel

We derive an input/output model in this section. We first show that the multipath effects can be modeled as a linear time-varying system. We then obtain a baseband representation of this model. The continuous-time channel is then sampled to obtain a discrete-time model. Finally we incorporate additive noise.

2.2.1 The wireless channel as a linear time-varying system

In the previous section we focused on the response to the sinusoidal input $\phi(t) = \cos 2\pi ft$. The received signal can be written as $\sum_i a_i(f, t)\phi(t - \tau_i(f, t))$, where $a_i(f, t)$ and $\tau_i(f, t)$ are respectively the overall attenuation and propagation delay at time t from the transmitter to the receiver on path i. The overall attenuation is simply the product of the attenuation factors due to the antenna pattern of the transmitter and the receiver, the nature of the reflector, as well as a factor that is a function of the distance from the transmitting antenna to the reflector and from the reflector to the receive antenna. We have described the channel effect at a particular frequency f. If we further assume that the $a_i(f, t)$ and the $\tau_i(f, t)$ do not depend on the frequency f, then we can use the principle of superposition to generalize the above input/output relation to an arbitrary input $x(t)$ with non-zero bandwidth:

$$y(t) = \sum_i a_i(t)x(t - \tau_i(t)). \tag{2.14}$$

In practice the attenuations and the propagation delays are usually slowly varying functions of frequency. These variations follow from the time-varying path lengths and also from frequency-dependent antenna gains. However, we are primarily interested in transmitting over bands that are narrow relative to the carrier frequency, and over such ranges we can omit this frequency dependence. It should however be noted that although the *individual* attenuations and delays are assumed to be independent of the frequency, the *overall* channel response can still vary with frequency due to the fact that different paths have different delays.

For the example of a perfectly reflecting wall in Figure 2.4, then,

$$a_1(t) = \frac{|\alpha|}{r_0 + vt}, \qquad a_2(t) = \frac{|\alpha|}{2d - r_0 - vt}, \tag{2.15}$$

$$\tau_1(t) = \frac{r_0 + vt}{c} - \frac{\angle\phi_1}{2\pi f}, \qquad \tau_2(t) = \frac{2d - r_0 - vt}{c} - \frac{\angle\phi_2}{2\pi f}, \tag{2.16}$$

where the first expression is for the direct path and the second for the reflected path. The term $\angle\phi_j$ here is to account for possible phase changes at the transmitter, reflector, and receiver. For the example here, there is a phase reversal at the reflector so we take $\phi_1 = 0$ and $\phi_2 = \pi$.

Since the channel (2.14) is linear, it can be described by the response $h(\tau, t)$ at time t to an impulse transmitted at time $t - \tau$. In terms of $h(\tau, t)$, the input/output relationship is given by

$$y(t) = \int_{-\infty}^{\infty} h(\tau, t) x(t - \tau)\, \mathrm{d}\tau. \tag{2.17}$$

Comparing (2.17) and (2.14), we see that the impulse response for the fading multipath channel is

$$h(\tau, t) = \sum_i a_i(t) \delta(\tau - \tau_i(t)). \tag{2.18}$$

This expression is really quite nice. It says that the effect of mobile users, arbitrarily moving reflectors and absorbers, and all of the complexities of solving Maxwell's equations, finally reduce to an input/output relation between transmit and receive antennas which is simply represented as the impulse response of a linear time-varying channel filter.

The effect of the Doppler shift is not immediately evident in this representation. From (2.16) for the single reflecting wall example, $\tau_i'(t) = v_i/c$ where v_i is the velocity with which the ith path length is increasing. Thus, the Doppler shift on the ith path is $-f\tau_i'(t)$.

In the special case when the transmitter, receiver and the environment are all stationary, the attenuations $a_i(t)$ and propagation delays $\tau_i(t)$ do not depend on time t, and we have the usual linear time-invariant channel with an impulse response

$$h(\tau) = \sum_i a_i \delta(\tau - \tau_i). \tag{2.19}$$

For the time-varying impulse response $h(\tau, t)$, we can define a time-varying frequency response

$$H(f; t) := \int_{-\infty}^{\infty} h(\tau, t) e^{-j2\pi f\tau} \, d\tau = \sum_i a_i(t) e^{-j2\pi f \tau_i(t)}. \tag{2.20}$$

In the special case when the channel is time-invariant, this reduces to the usual frequency response. One way of interpreting $H(f; t)$ is to think of the system as a slowly varying function of t with a frequency response $H(f; t)$ at each fixed time t. Corresponding, $h(\tau, t)$ can be thought of as the impulse response of the system at a fixed time t. This is a legitimate and useful way of thinking about many multipath fading channels, as the time-scale at which the channel varies is typically much longer than the delay spread (i.e., the amount of memory) of the impulse response at a fixed time. In the reflecting wall example in Section 2.1.4, the time taken for the channel to change significantly is of the order of milliseconds while the delay spread is of the order of microseconds. Fading channels which have this characteristic are sometimes called *underspread* channels.

2.2.2 Baseband equivalent model

In typical wireless applications, communication occurs in a passband $[f_c - W/2, f_c + W/2]$ of bandwidth W around a center frequency f_c, the spectrum having been specified by regulatory authorities. However, most of the processing, such as coding/decoding, modulation/demodulation, synchronization, etc., is actually done at the baseband. At the transmitter, the last stage of the operation is to "up-convert" the signal to the carrier frequency and transmit it via the antenna. Similarly, the first step at the receiver is to "down-convert" the RF (radio-frequency) signal to the baseband before further processing. Therefore from a communication system design point of view, it is most useful to have a baseband equivalent representation of the system. We first start with defining the baseband equivalent representation of signals.

Consider a real signal $s(t)$ with Fourier transform $S(f)$, band-limited in $[f_c - W/2, f_c + W/2]$ with $W < 2f_c$. Define its *complex baseband equivalent* $s_b(t)$ as the signal having Fourier transform:

$$S_b(f) = \begin{cases} \sqrt{2} S(f + f_c) & f + f_c > 0, \\ 0 & f + f_c \le 0. \end{cases} \tag{2.21}$$

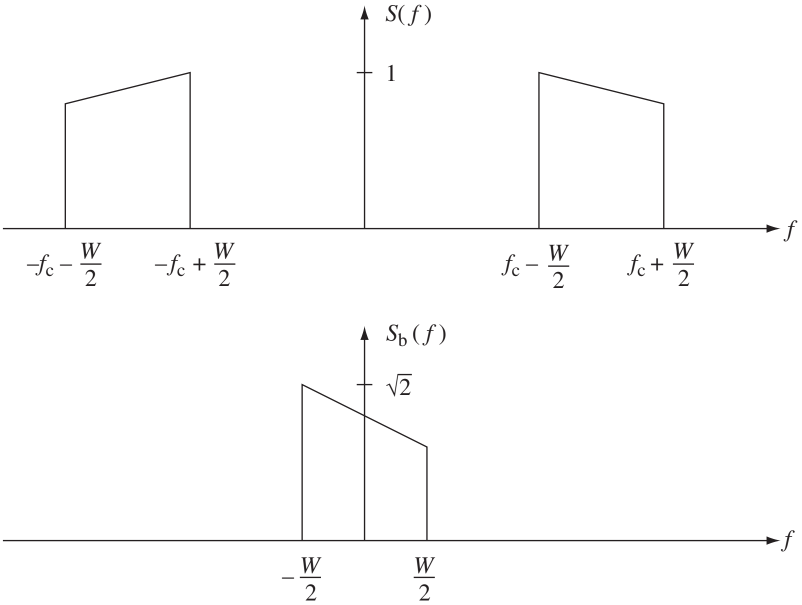


Figure 2.7 Illustration of the relationship between a passband spectrum *S(f)* and its baseband equivalent $S_b(f)$.

Since $s(t)$ is real, its Fourier transform satisfies $S(f) = S^*(-f)$, which means that $s_b(t)$ contains exactly the same information as $s(t)$. The factor of $\sqrt{2}$ is quite arbitrary but chosen to normalize the energies of $s_b(t)$ and $s(t)$ to be the same. Note that $s_b(t)$ is band-limited in $[-W/2, W/2]$. See Figure 2.7.

To reconstruct $s(t)$ from $s_b(t)$, we observe that

$$\sqrt{2}S(f) = S_b(f - f_c) + S_b^*(-f - f_c). \tag{2.22}$$

Taking inverse Fourier transforms, we get

$$s(t) = \frac{1}{\sqrt{2}}\left[s_b(t)e^{j2\pi f_c t} + s_b^*(t)e^{-j2\pi f_c t}\right] = \sqrt{2}\Re\left[s_b(t)e^{j2\pi f_c t}\right]. \tag{2.23}$$

In terms of real signals, the relationship between $s(t)$ and $s_b(t)$ is shown in Figure 2.8. The passband signal $s(t)$ is obtained by modulating $\Re[s_b(t)]$ by $\sqrt{2}\cos 2\pi f_c t$ and $\Im[s_b(t)]$ by $-\sqrt{2}\sin 2\pi f_c t$ and summing, to get $\sqrt{2}\Re\left[s_b(t)e^{j2\pi f_c t}\right]$ (up-conversion). The baseband signal $\Re[s_b(t)]$ (respectively $\Im[s_b(t)]$) is obtained by modulating $s(t)$ by $\sqrt{2}\cos 2\pi f_c t$ (respectively $-\sqrt{2}\sin 2\pi f_c t$) followed by ideal low-pass filtering at the baseband $[-W/2, W/2]$ (down-conversion).

Let us now go back to the multipath fading channel (2.14) with impulse response given by (2.18). Let $x_b(t)$ and $y_b(t)$ be the complex baseband equivalents of the transmitted signal $x(t)$ and the received signal $y(t)$, respectively. Figure 2.9 shows the system diagram from $x_b(t)$ to $y_b(t)$. This implementation of a passband communication system is known as *quadrature amplitude modulation* (QAM). The signal $\Re[x_b(t)]$ is sometimes called the

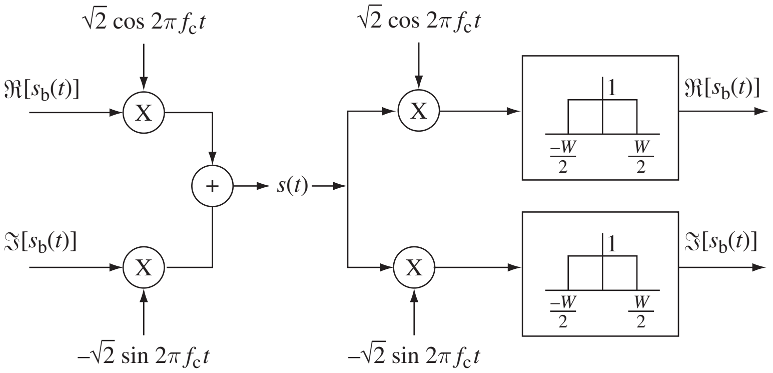


Figure 2.8 Illustration of upconversion from $s_b(t)$ to $s(t)$, followed by downconversion from $s(t)$ back to $s_b(t)$.

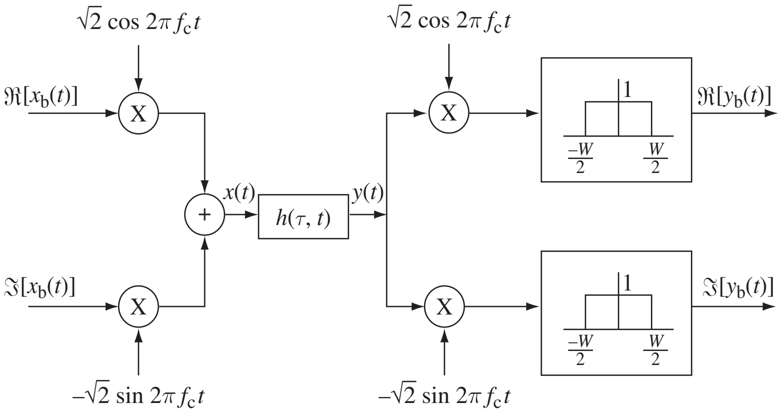


Figure 2.9 System diagram from the baseband transmitted signal $x_b(t)$ to the baseband received signal $y_b(t)$.

in-phase component I and $\Im[x_b(t)]$ the quadrature component Q (rotated by $\pi/2$). We now calculate the baseband equivalent channel. Substituting $x(t) = \sqrt{2}\Re[x_b(t)e^{j2\pi f_c t}]$ and $y(t) = \sqrt{2}\Re[y_b(t)e^{j2\pi f_c t}]$ into (2.14) we get

$$\begin{aligned}\Re[y_b(t)e^{j2\pi f_c t}] &= \sum_i a_i(t)\Re[x_b(t-\tau_i(t))e^{j2\pi f_c(t-\tau_i(t))}]\\ &= \Re\left[\left\{\sum_i a_i(t)x_b(t-\tau_i(t))e^{-j2\pi f_c \tau_i(t)}\right\}e^{j2\pi f_c t}\right]. \qquad (2.24)\end{aligned}$$

Similarly, one can obtain (Exercise 2.13)

$$\Im[y_b(t)e^{j2\pi f_c t}] = \Im\left[\left\{\sum_i a_i(t)x_b(t-\tau_i(t))e^{-j2\pi f_c \tau_i(t)}\right\}e^{j2\pi f_c t}\right]. \qquad (2.25)$$

Hence, the baseband equivalent channel is

$$y_b(t) = \sum_i a_i^b(t)x_b(t-\tau_i(t)), \qquad (2.26)$$

where

$$a_i^{\mathrm{b}}(t) := a_i(t)\mathrm{e}^{-\mathrm{j}2\pi f_{\mathrm{c}}\tau_i(t)}. \tag{2.27}$$

The input/output relationship in (2.26) is also that of a linear time-varying system, and the baseband equivalent impulse response is

$$h_{\mathrm{b}}(\tau, t) = \sum_i a_i^{\mathrm{b}}(t)\delta(\tau - \tau_i(t)). \tag{2.28}$$

This representation is easy to interpret in the time domain, where the effect of the carrier frequency can be seen explicitly. The baseband output is the sum, over each path, of the delayed replicas of the baseband input. The magnitude of the ith such term is the magnitude of the response on the given path; this changes slowly, with significant changes occurring on the order of seconds or more. The phase is changed by $\pi/2$ (i.e., is changed significantly) when the delay on the path changes by $1/(4f_{\mathrm{c}})$, or equivalently, when the path length changes by a quarter wavelength, i.e., by $c/(4f_{\mathrm{c}})$. If the path length is changing at velocity v, the time required for such a phase change is $c/(4f_{\mathrm{c}}v)$. Recalling that the Doppler shift D at frequency f is fv/c, and noting that $f \approx f_{\mathrm{c}}$ for narrowband communication, the time required for a $\pi/2$ phase change is $1/(4D)$. For the single reflecting wall example, this is about 5 ms (assuming $f_{\mathrm{c}} = 900\,\mathrm{MHz}$ and $v = 60\,\mathrm{km/h}$). The phases of both paths are rotating at this rate but in opposite directions.

Note that the Fourier transform $H_{\mathrm{b}}(f; t)$ of $h_{\mathrm{b}}(\tau, t)$ for a fixed t is simply $H(f + f_{\mathrm{c}}; t)$, i.e., the frequency response of the original system (at a fixed t) shifted by the carrier frequency. This provides another way of thinking about the baseband equivalent channel.

2.2.3 A discrete-time baseband model

The next step in creating a useful channel model is to convert the continuous-time channel to a discrete-time channel. We take the usual approach of the sampling theorem. Assume that the input waveform is band-limited to W. The baseband equivalent is then limited to $W/2$ and can be represented as

$$x_{\mathrm{b}}(t) = \sum_n x[n]\mathrm{sinc}(Wt - n), \tag{2.29}$$

where $x[n]$ is given by $x_{\mathrm{b}}(n/W)$ and $\mathrm{sinc}(t)$ is defined as

$$\mathrm{sinc}(t) := \frac{\sin(\pi t)}{\pi t}. \tag{2.30}$$

This representation follows from the sampling theorem, which says that any waveform band-limited to $W/2$ can be expanded in terms of the orthogonal basis $\{\text{sinc}(Wt-n)\}_n$, with coefficients given by the samples (taken uniformly at integer multiples of $1/W$).

Using (2.26), the baseband output is given by

$$y_\text{b}(t) = \sum_n x[n] \sum_i a_i^\text{b}(t)\text{sinc}(Wt - W\tau_i(t) - n). \tag{2.31}$$

The sampled outputs at multiples of $1/W$, $y[m] := y_\text{b}(m/W)$, are then given by

$$y[m] = \sum_n x[n] \sum_i a_i^\text{b}(m/W)\text{sinc}[m - n - \tau_i(m/W)W]. \tag{2.32}$$

The sampled output $y[m]$ can equivalently be thought of as the projection of the waveform $y_\text{b}(t)$ onto the waveform $W\text{sinc}(Wt-m)$. Let $\ell := m-n$. Then

$$y[m] = \sum_\ell x[m-\ell] \sum_i a_i^\text{b}(m/W)\text{sinc}[\ell - \tau_i(m/W)W]. \tag{2.33}$$

By defining

$$\boxed{h_\ell[m] := \sum_i a_i^\text{b}(m/W)\text{sinc}[\ell - \tau_i(m/W)W],} \tag{2.34}$$

(2.33) can be written in the simple form

$$y[m] = \sum_\ell h_\ell[m]\, x[m-\ell]. \tag{2.35}$$

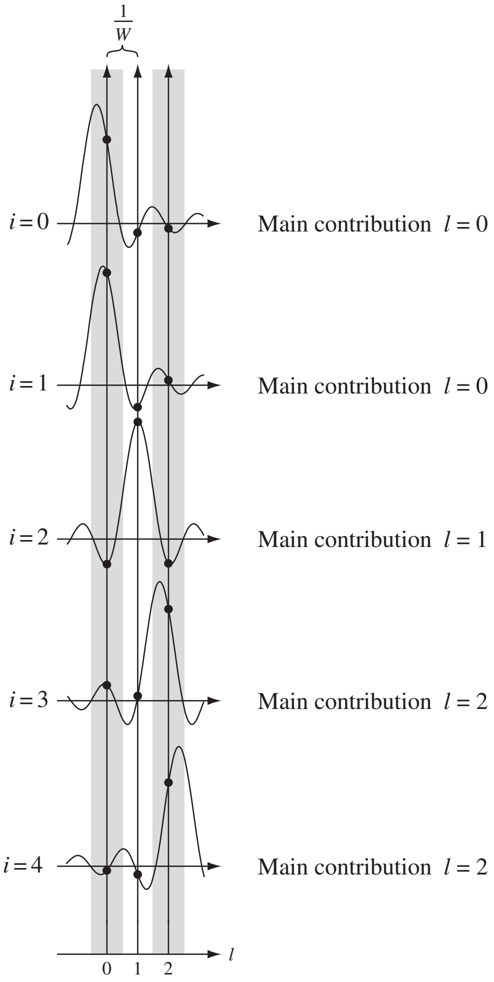
We denote $h_\ell[m]$ as the ℓth (complex) channel filter tap at time m. Its value is a function of mainly the gains $a_i^\text{b}(t)$ of the paths, whose delays $\tau_i(t)$ are close to ℓ/W (Figure 2.10). In the special case where the gains $a_i^\text{b}(t)$ and the delays $\tau_i(t)$ of the paths are time-invariant, (2.34) simplifies to

$$h_\ell = \sum_i a_i^\text{b}\ \text{sinc}[\ell - \tau_i W], \tag{2.36}$$

and the channel is linear time-invariant. The ℓth tap can be interpreted as the sample (ℓ/W)th of the low-pass filtered baseband channel response $h_\text{b}(\tau)$ (cf. (2.19)) convolved with $\text{sinc}(W\tau)$.

We can interpret the sampling operation as modulation and demodulation in a communication system. At time n, we are modulating the complex symbol $x[m]$ (in-phase plus quadrature components) by the sinc pulse before the up-conversion. At the receiver, the received signal is sampled at times m/W

Figure 2.10 Due to the decay of the sinc function, the ith path contributes most significantly to the ℓth tap if its delay falls in the window $[\ell/W - 1/(2W), \ell/W + 1/(2W)]$.



$\frac{1}{W}$
$i = 0$
Main contribution $l = 0$
$i = 1$
Main contribution $l = 0$
$i = 2$
Main contribution $l = 1$
$i = 3$
Main contribution $l = 2$
$i = 4$
Main contribution $l = 2$
l
0
1
2

at the output of the low-pass filter. Figure 2.11 shows the complete system. In practice, other transmit pulses, such as the raised cosine pulse, are often used in place of the sinc pulse, which has rather poor time-decay property and tends to be more susceptible to timing errors. This necessitates sampling at the Nyquist sampling rate, but does not alter the essential nature of the model. Hence we will confine to Nyquist sampling.

Due to the Doppler spread, the bandwidth of the output $y_b(t)$ is generally slightly larger than the bandwidth $W/2$ of the input $x_b(t)$, and thus the output samples $\{y[m]\}$ do not fully represent the output waveform. This problem is usually ignored in practice, since the Doppler spread is small (of the order of tens to hundreds of Hz) compared to the bandwidth W. Also, it is very convenient for the sampling rate of the input and output to be the same. Alternatively, it would be possible to sample the output at twice the rate of the input. This would recapture all the information in the received waveform.

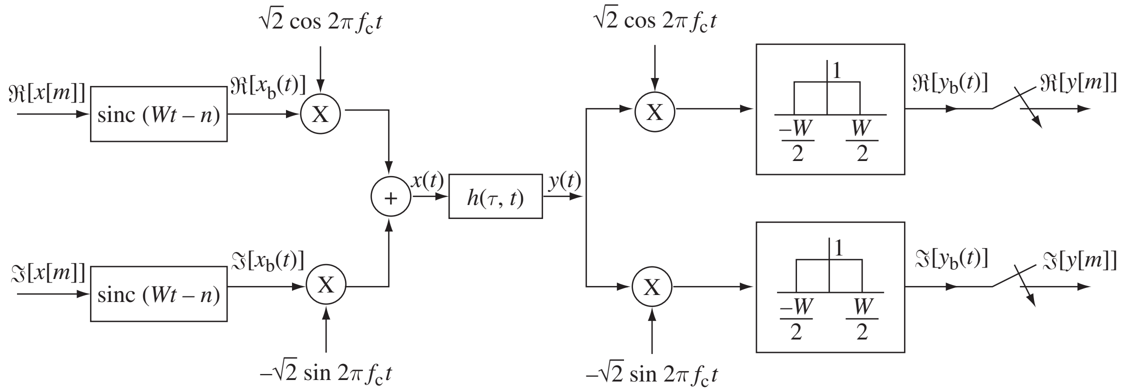


Figure 2.11 System diagram from the baseband transmitted symbol $x[m]$ to the baseband sampled received signal $y[m]$.

The number of taps would be almost doubled because of the reduced sample interval, but it would typically be somewhat less than doubled since the representation would not spread the path delays so much.

Discussion 2.1 Degrees of freedom

The symbol $x[m]$ is the mth sample of the transmitted signal; there are W samples per second. Each symbol is a complex number; we say that it represents one (complex) *dimension* or *degree of freedom*. The continuous-time signal $x(t)$ of duration one second corresponds to W discrete symbols; thus we could say that the band-limited, continuous-time signal has W degrees of freedom, per second.

The mathematical justification for this interpretation comes from the following important result in communication theory: the signal space of complex continuous-time signals of duration T which have most of their energy within the frequency band $[-W/2, W/2]$ has dimension approximately WT. (A precise statement of this result is in standard communication theory text/books; see Section 5.3 of [148] for example.) This result reinforces our interpretation that a continuous-time signal with bandwidth W can be represented by W complex dimensions per second.

The received signal $y(t)$ is also band-limited to approximately W (due to the Doppler spread, the bandwidth is slightly larger than W) and has W complex dimensions per second. From the point of view of communication over the channel, the *received* signal space is what matters because it dictates the number of different signals which can be reliably distinguished at the receiver. Thus, we define the *degrees of freedom of the channel* to be the dimension of the received signal space, and whenever we refer to the signal space, we implicitly mean the received signal space unless stated otherwise.

2.2.4 Additive white noise

As a last step, we include additive noise in our input/output model. We make the standard assumption that $w(t)$ is zero-mean additive white Gaussian noise (AWGN) with power spectral density $N_0/2$ (i.e., $E[w(0)w(t)] = (N_0/2)\delta(t)$. The model (2.14) is now modified to be

$$y(t) = \sum_i a_i(t)x(t - \tau_i(t)) + w(t). \tag{2.37}$$

See Figure 2.12. The discrete-time baseband-equivalent model (2.35) now becomes

$$y[m] = \sum_\ell h_\ell[m]x[m - \ell] + w[m], \tag{2.38}$$

where $w[m]$ is the low-pass filtered noise at the sampling instant m/W. Just like the signal, the white noise $w(t)$ is down-converted, filtered at the baseband and ideally sampled. Thus, it can be verified (Exercise 2.11) that

$$\Re(w[m]) = \int_{-\infty}^{\infty} w(t)\psi_{m,1}(t)\mathrm{d}t, \tag{2.39}$$

$$\Im(w[m]) = \int_{-\infty}^{\infty} w(t)\psi_{m,2}(t)\mathrm{d}t, \tag{2.40}$$

where

$$\begin{aligned} \psi_{m,1}(t) &:= \sqrt{2W}\cos(2\pi f_c t)\mathrm{sinc}(Wt - m), \\ \psi_{m,2}(t) &:= -\sqrt{2W}\sin(2\pi f_c t)\mathrm{sinc}(Wt - m). \end{aligned} \tag{2.41}$$

It can further be shown that $\{\psi_{m,1}(t), \psi_{m,2}(t)\}_m$ forms an *orthonormal set* of waveforms, i.e., the waveforms are orthogonal to each other (Exercise 2.12). In Appendix A we review the definition and basic properties of white Gaussian random *vectors* (i.e., vectors whose components are independent and identically distributed (i.i.d.) Gaussian random variables). A key property is that the projections of a white Gaussian random vector onto any orthonormal vectors are independent and identically distributed Gaussian random variables. Heuristically, one can think of continuous-time Gaussian white noise as an infinite-dimensional white random vector and the above property carries through: the projections onto orthogonal waveforms are uncorrelated and hence independent. Hence the discrete-time noise process $\{w[m]\}$ is white, i.e., independent over time; moreover, the real and imaginary components are i.i.d. Gaussians with variances $N_0/2$. A complex Gaussian random variable X whose real and imaginary components are i.i.d. satisfies a *circular symmetry* property: $\mathrm{e}^{\mathrm{j}\phi}X$ has the same distribution as X for any ϕ. We shall call such a random variable *circular symmetric complex*

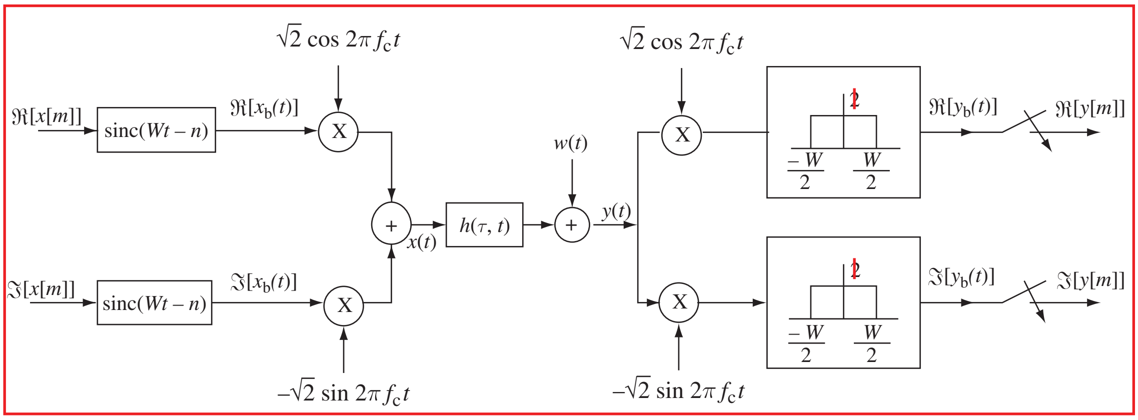


Figure 2.12 A complete system diagram.

Gaussian, denoted by $\mathcal{CN}(0,\sigma^2)$, where $\sigma^2 = E[|X|^2]$. The concept of circular symmetry is discussed further in Section A.1.3 of Appendix A.

The assumption of AWGN essentially means that we are assuming that the primary source of the noise is at the receiver or is radiation impinging on the receiver that is independent of the paths over which the signal is being received. This is normally a very good assumption for most communication situations.

2.3 Time and frequency coherence

2.3.1 Doppler spread and coherence time

An important channel parameter is the time-scale of the variation of the channel. How fast do the taps $h_\ell[m]$ vary as a function of time m? Recall that

$$\begin{aligned} h_\ell[m] &= \sum_i a_i^{\mathrm{b}}(m/W)\mathrm{sinc}[\ell-\tau_i(m/W)W] \\ &= \sum_i a_i(m/W)\mathrm{e}^{-\mathrm{j}2\pi f_c\tau_i(m/W)}\mathrm{sinc}[\ell-\tau_i(m/W)W]. \qquad (2.42) \end{aligned}$$

Let us look at this expression term by term. From Section 2.2.2 we gather that significant changes in a_i occur over periods of seconds or more. Significant changes in the phase of the ith path occur at intervals of $1/(4D_i)$, where $D_i = f_c\tau_i'(t)$ is the Doppler shift for that path. When the different paths contributing to the ℓth tap have different Doppler shifts, the magnitude of $h_\ell[m]$ changes significantly. This is happening at the time-scale inversely proportional to the largest difference between the Doppler shifts, the *Doppler spread* D_s:

$$\boxed{D_s := \max_{i,j} f_c|\tau_i'(t)-\tau_j'(t)|,} \qquad (2.43)$$

where the maximum is taken over all the paths that contribute significantly to a tap.[7] Typical intervals for such changes are on the order of 10 ms. Finally, changes in the sinc term of (2.42) due to the time variation of each $\tau_i(t)$ are proportional to the bandwidth, whereas those in the phase are proportional to the carrier frequency, which is typically much larger. Essentially, it takes much longer for a path to move from one tap to the next than for its phase to change significantly. Thus, the fastest changes in the filter taps occur because of the phase changes, and these are significant over delay changes of $1/(4D_\mathrm{s})$.

The coherence time T_c of a wireless channel is defined (in an order of magnitude sense) as the interval over which $h_\ell[m]$ changes significantly as a function of m. What we have found, then, is the important relation

$$\boxed{T_\mathrm{c} = \frac{1}{4D_\mathrm{s}}.} \tag{2.44}$$

This is a somewhat imprecise relation, since the largest Doppler shifts may belong to paths that are too weak to make a difference. We could also view a phase change of $\pi/4$ to be significant, and thus replace the factor of 4 above by 8. Many people instead replace the factor of 4 by 1. The important thing is to recognize that the major effect in determining time coherence is the Doppler spread, and that the relationship is reciprocal; the larger the Doppler spread, the smaller the time coherence.

In the wireless communication literature, channels are often categorized as *fast fading* and *slow fading*, but there is little consensus on what these terms mean. In this book, we will call a channel fast fading if the coherence time T_c is much shorter than the delay requirement of the application, and slow fading if T_c is longer. The operational significance of this definition is that, in a fast fading channel, one can transmit the coded symbols over multiple fades of the channel, while in a slow fading channel, one cannot. Thus, whether a channel is fast or slow fading depends not only on the environment but also on the application; voice, for example, typically has a short delay requirement of less than 100 ms, while some types of data applications can have a laxer delay requirement.

2.3.2 Delay spread and coherence bandwidth

Another important general parameter of a wireless system is the multipath delay spread, T_d, defined as the difference in propagation time between the

[7] The Doppler spread can in principle be different for different taps. Exercise 2.10 explores this possibility.

longest and shortest path, counting only the paths with significant energy. Thus,

$$\boxed{T_{\mathrm{d}} := \max_{i,j} |\tau_i(t) - \tau_j(t)|.} \tag{2.45}$$

This is defined as a function of t, but we regard it as an order of magnitude quantity, like the time coherence and Doppler spread. If a cell or LAN has a linear extent of a few kilometers or less, it is very unlikely to have path lengths that differ by more than 300 to 600 meters. This corresponds to path delays of one or two microseconds. As cells become smaller due to increased cellular usage, T_{d} also shrinks. As was already mentioned, typical wireless channels are underspread, which means that the delay spread T_{d} is much smaller than the coherence time T_{c}.

The bandwidths of cellular systems range between several hundred kilohertz and several megahertz, and thus, for the above multipath delay spread values, all the path delays in (2.34) lie within the peaks of two or three sinc functions; more often, they lie within a single peak. Adding a few extra taps to each channel filter because of the slow decay of the sinc function, we see that cellular channels can be represented with at most four or five channel filter taps. On the other hand, there is a recent interest in *ultra-wideband* (UWB) communication, operating from 3.1 to 10.6 GHz. These channels can have up to a few hundred taps.

When we study modulation and detection for cellular systems, we shall see that the receiver must estimate the values of these channel filter taps. The taps are estimated via transmitted and received waveforms, and thus the receiver makes no explicit use of (and usually does not have) any information about individual path delays and path strengths. This is why we have not studied the details of propagation over multiple paths with complicated types of reflection mechanisms. All we really need is the aggregate values of gross physical mechanisms such as Doppler spread, coherence time, and multipath spread.

The delay spread of the channel dictates its *frequency coherence*. Wireless channels change both in time and frequency. The time coherence shows us how quickly the channel changes in time, and similarly, the frequency coherence shows how quickly it changes in frequency. We first understood about channels changing in time, and correspondingly about the duration of fades, by studying the simple example of a direct path and a single reflected path. That same example also showed us how channels change with frequency. We can see this in terms of the frequency response as well.

Recall that the frequency response at time t is

$$H(f; t) = \sum_i a_i(t) e^{-j2\pi f \tau_i(t)}. \tag{2.46}$$

The contribution due to a particular path has a phase linear in f. For multiple paths, there is a differential phase, $2\pi f(\tau_i(t) - \tau_k(t))$. This differential

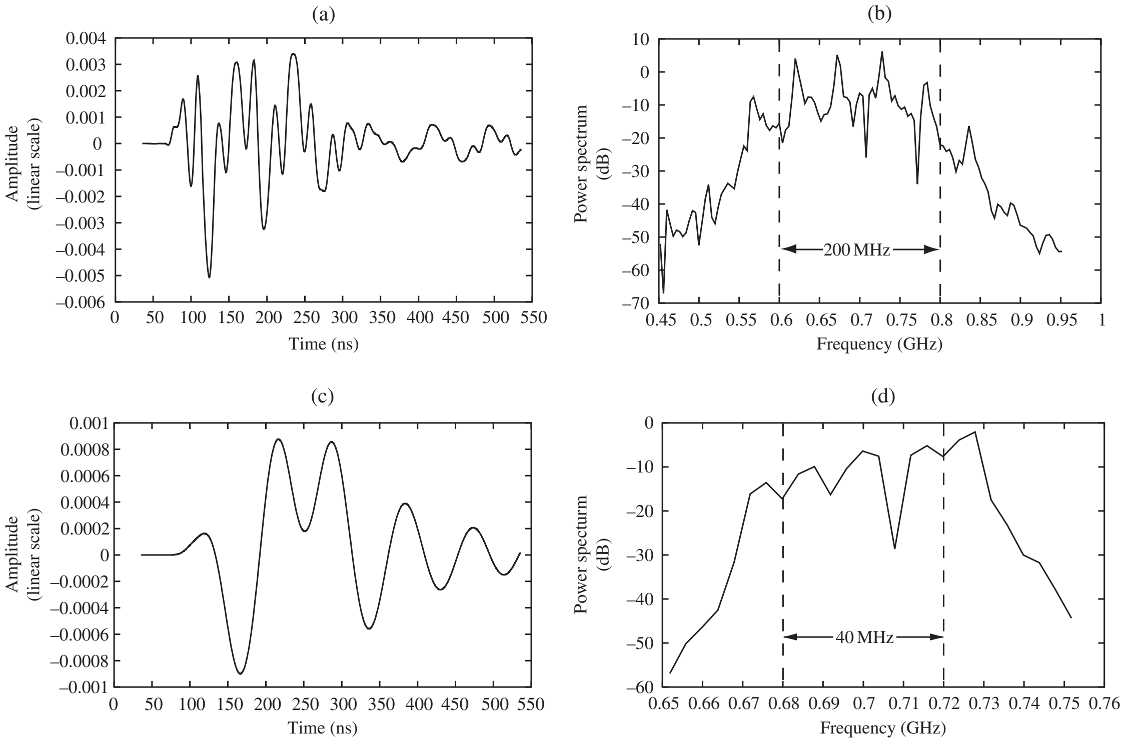


Figure 2.13 (a) A channel over 200 MHz is frequency-selective, and the impulse response has many taps. (b) The spectral content of the same channel. (c) The same channel over 40 MHz is flatter, and has for fewer taps. (d) The spectral contents of the same channel, limited to 40 MHz bandwidth. At larger bandwidths, the same physical paths are resolved into a finer resolution.

phase causes selective fading in frequency. This says that $E_{\mathrm{r}}(f, t)$ changes significantly, not only when t changes by $1/(4D_{\mathrm{s}})$, but also when f changes by $1/(2T_{\mathrm{d}})$. This argument extends to an arbitrary number of paths, so the *coherence bandwidth*, W_{c}, is given by

$$\boxed{W_{\mathrm{c}} = \frac{1}{2T_{\mathrm{d}}}.} \tag{2.47}$$

This relationship, like (2.44), is intended as an order of magnitude relation, essentially pointing out that the coherence bandwidth is reciprocal to the multipath spread. When the bandwidth of the input is considerably less than W_{c}, the channel is usually referred to as *flat fading*. In this case, the delay spread T_{d} is much less than the symbol time $1/W$, and a single channel filter tap is sufficient to represent the channel. When the bandwidth is much larger than W_{c}, the channel is said to be *frequency-selective*, and it has to be represented by multiple taps. Note that flat or frequency-selective fading is not a property of the channel alone, but of the relationship between the bandwidth W and the coherence bandwidth T_{d} (Figure 2.13).

The physical parameters and the time-scale of change of key parameters of the discrete-time baseband channel model are summarized in Table 2.1. The different types of channels are summarized in Table 2.2.

Table 2.1 A summary of the physical parameters of the channel and the time-scale of change of the key parameters in its discrete-time baseband model.

Key channel parameters and time-scales	Symbol	Representative values
Carrier frequency	f_c	1 GHz
Communication bandwidth	W	1 MHz
Distance between transmitter and receiver	d	1 km
Velocity of mobile	v	64 km/h
Doppler shift for a path	$D = f_c v/c$	50 Hz
Doppler spread of paths corresponding to a tap	D_s	100 Hz
Time-scale for change of path amplitude	d/v	1 minute
Time-scale for change of path phase	$1/(4D)$	5 ms
Time-scale for a path to move over a tap	$c/(vW)$	20 s
Coherence time	$T_c = 1/(4D_s)$	2.5 ms
Delay spread	T_d	1 μs
Coherence bandwidth	$W_c = 1/(2T_d)$	500 kHz

Table 2.2 A summary of the types of wireless channels and their defining characteristics.

Types of channel	Defining characteristic
Fast fading	$T_c \ll$ delay requirement
Slow fading	$T_c \gg$ delay requirement
Flat fading	$W \ll W_c$
Frequency-selective fading	$W \gg W_c$
Underspread	$T_d \ll T_c$

2.4 Statistical channel models

2.4.1 Modeling philosophy

We defined Doppler spread and multipath spread in the previous section as quantities associated with a given receiver at a given location, velocity, and time. However, we are interested in a characterization that is valid over some range of conditions. That is, we recognize that the channel filter taps $\{h_\ell[m]\}$ must be measured, but we want a statistical characterization of how many taps are necessary, how quickly they change and how much they vary.

Such a characterization requires a probabilistic model of the channel tap values, perhaps gathered by statistical measurements of the channel. We are familiar with describing additive noise by such a probabilistic model (as a Gaussian random variable). We are also familiar with evaluating error probability while communicating over a channel using such models. These error probability evaluations, however, depend critically on the independence and Gaussian distribution of the noise variables.

It should be clear from the description of the physical mechanisms generating Doppler spread and multipath spread that probabilistic models for the channel filter taps are going to be far less believable than the models for additive noise. On the other hand, we need such models, even if they are quite inaccurate. Without models, systems are designed using experience and experimentation, and creativity becomes somewhat stifled. Even with highly over-simplified models, we can compare different system approaches and get a sense of what types of approaches are worth pursuing.

To a certain extent, all analytical work is done with simplified models. For example, white Gaussian noise (WGN) is often assumed in communication models, although we know the model is valid only over sufficiently small frequency bands. With WGN, however, we expect the model to be quite good when used properly. For wireless channel models, however, probabilistic models are quite poor and only provide order-of-magnitude guides to system design and performance. We will see that we can define Doppler spread, multipath spread, etc. much more cleanly with probabilistic models, but the underlying problem remains that these channels are very different from each other and cannot really be characterized by probabilistic models. At the same time, there is a large literature based on probabilistic models for wireless channels, and it has been highly useful for providing insight into wireless systems. However, it is important to understand the robustness of results based on these models.

There is another question in deciding what to model. Recall the continuous-time multipath fading channel

$$y(t) = \sum_i a_i(t) x(t - \tau_i(t)) + w(t). \tag{2.48}$$

This contains an exact specification of the delay and magnitude of each path. From this, we derived a discrete-time baseband model in terms of channel filter taps as

$$y[m] = \sum_\ell h_\ell[m] x[m - \ell] + w[m], \tag{2.49}$$

where

$$h_\ell[m] = \sum_i a_i(m/W) \mathrm{e}^{-\mathrm{j}2\pi f_c \tau_i(m/W)} \mathrm{sinc}[\ell - \tau_i(m/W)W]. \tag{2.50}$$

We used the sampling theorem expansion in which $x[m] = x_\mathrm{b}(m/W)$ and $y[m] = y_\mathrm{b}(m/W)$. Each channel tap $h_\ell[m]$ contains an aggregate of paths, with the delays smoothed out by the baseband signal bandwidth.

Fortunately, it is the filter taps that must be modeled for input/output descriptions, and also fortunately, the filter taps often contain a sufficient path aggregation so that a statistical model might have a chance of success.

2.4.2 Rayleigh and Rician fading

The simplest probabilistic model for the channel filter taps is based on the assumption that there are a large number of statistically independent reflected and scattered paths with random amplitudes in the delay window corresponding to a single tap. The phase of the ith path is $2\pi f_c \tau_i$ modulo 2π. Now, $f_c \tau_i = d_i/\lambda$, where d_i is the distance travelled by the ith path and λ is the carrier wavelength. Since the reflectors and scatterers are far away relative to the carrier wavelength, i.e., $d_i \gg \lambda$, it is reasonable to assume that the phase for each path is uniformly distributed between 0 and 2π and that the phases of different paths are independent. The contribution of each path in the tap gain $h_\ell[m]$ is

$$a_i(m/W)\mathrm{e}^{-\mathrm{j}2\pi f_c \tau_i(m/W)}\operatorname{sinc}[\ell - \tau_i(m/W)W] \qquad (2.51)$$

and this can be modeled as a circular symmetric complex random variable.[8] Each tap $h_\ell[m]$ is the sum of a large number of such small independent circular symmetric random variables. It follows that $\Re(h_\ell[m])$ is the sum of many small independent real random variables, and so by the Central Limit Theorem, it can reasonably be modeled as a zero-mean Gaussian random variable. Similarly, because of the uniform phase, $\Re(h_\ell[m]\mathrm{e}^{\mathrm{j}\phi})$ is Gaussian with the same variance for any fixed ϕ. This assures us that $h_\ell[m]$ is in fact circular symmetric $\mathcal{CN}(0, \sigma_\ell^2)$ (see Section A.1.3 in Appendix A for an elaboration). It is assumed here that the variance of $h_\ell[m]$ is a function of the tap ℓ, but independent of time m (there is little point in creating a probabilistic model that depends on time). With this assumed Gaussian probability density, we know that the magnitude $|h_\ell[m]|$ of the ℓth tap is a *Rayleigh* random variable with density (cf. (A.20) in Appendix A and Exercise 2.14)

$$\boxed{\frac{x}{\sigma_\ell^2}\exp\left\{\frac{-x^2}{2\sigma_\ell^2}\right\}, \quad x \geq 0,} \qquad (2.52)$$

and the squared magnitude $|h_\ell[m]|^2$ is exponentially distributed with density

$$\boxed{\frac{1}{\sigma_\ell^2}\exp\left\{\frac{-x}{\sigma_\ell^2}\right\}, \quad x \geq 0.} \qquad (2.53)$$

This model, which is called *Rayleigh fading*, is quite reasonable for scattering mechanisms where there are many small reflectors, but is adopted primarily for its simplicity in typical cellular situations with a relatively small number of reflectors. The word *Rayleigh* is almost universally used for this

[8] See Section A.1.3 in Appendix A for a more in-depth discussion of circular symmetric random variables and vectors.

model, but the assumption is that the tap gains are circularly symmetric complex Gaussian random variables.

There is a frequently used alternative model in which the line-of-sight path (often called a *specular* path) is large and has a known magnitude, and that there are also a large number of independent paths. In this case, $h_\ell[m]$, at least for one value of ℓ, can be modeled as

$$\boxed{h_\ell[m] = \sqrt{\frac{\kappa}{\kappa+1}}\sigma_\ell e^{j\theta} + \sqrt{\frac{1}{\kappa+1}}\mathcal{CN}\left(0, \sigma_\ell^2\right)} \tag{2.54}$$

with the first term corresponding to the specular path arriving with uniform phase θ and the second term corresponding to the aggregation of the large number of reflected and scattered paths, independent of θ. The parameter κ (so-called K-factor) is the ratio of the energy in the specular path to the energy in the scattered paths; the larger κ is, the more deterministic is the channel. The magnitude of such a random variable is said to have a *Rician* distribution. Its density has quite a complicated form; it is often a better model of fading than the Rayleigh model.

2.4.3 Tap gain auto-correlation function

Modeling each $h_\ell[m]$ as a complex random variable provides part of the statistical description that we need, but this is not the most important part. The more important issue is how these quantities vary with time. As we will see in the rest of the book, the rate of channel variation has significant impact on several aspects of the communication problem. A statistical quantity that models this relationship is known as the *tap gain auto-correlation function*, $R_\ell[n]$. It is defined as

$$R_\ell[n] := \mathbb{E}\left\{h_\ell^*[m]h_\ell[m+n]\right\}. \tag{2.55}$$

For each tap ℓ, this gives the auto-correlation function of the sequence of random variables modeling that tap as it evolves in time. We are tacitly assuming that this is not a function of time m. Since the sequence of random variables $\{h_\ell[m]\}$ for any given ℓ has both a mean and covariance function that does not depend on m, this sequence is wide-sense stationary. We also assume that, as a random variable, $h_\ell[m]$ is independent of $h_{\ell'}[m']$ for all $\ell \neq \ell'$ and all m, m'. This final assumption is intuitively plausible since paths in different ranges of delay contribute to $h_\ell[m]$ for different values of ℓ.[9]

The coefficient $R_\ell[0]$ is proportional to the energy received in the ℓth tap. The multipath spread T_d can be defined as the product of $1/W$ times the range of ℓ which contains most of the total energy $\sum_{\ell=0}^{\infty} R_\ell[0]$. This is

[9] One could argue that a moving reflector would gradually travel from the range of one tap to another, but as we have seen, this typically happens over a very large time-scale.

somewhat preferable to our previous "definition" in that the statistical nature of T_d becomes explicit and the reliance on some sort of stationarity becomes explicit. Now, we can also define the coherence time T_c more explicitly as the smallest value of $n > 0$ for which $R_\ell[n]$ is significantly different from $R_\ell[0]$. With both of these definitions, we still have the ambiguity of what "significant" means, but we are now facing the reality that these quantities must be viewed as statistics rather than as instantaneous values.

The tap gain auto-correlation function is useful as a way of expressing the statistics for how tap gains change given a particular bandwidth W, but gives little insight into questions related to choice of a bandwidth for communication. If we visualize increasing the bandwidth, we can see several things happening. First, the ranges of delay that are separated into different taps ℓ become narrower ($1/W$ seconds), so there are fewer paths corresponding to each tap, and thus the Rayleigh approximation becomes poorer. Second, the sinc functions of (2.50) become narrower, and $R_\ell[0]$ gives a finer grained picture of the amount of power being received in the ℓth delay window of width $1/W$. In summary, as we try to apply this model to larger W, we get more detailed information about delay and correlation at that delay, but the information becomes more questionable.

Example 2.2 Clarke's model

This is a popular statistical model for flat fading. The transmitter is fixed, the mobile receiver is moving at speed v, and the transmitted signal is scattered by stationary objects around the mobile. There are K paths, the ith path arriving at an angle $\theta_i := 2\pi i/K$, $i = 0, \ldots, K-1$, with respect to the direction of motion. K is assumed to be large. The scattered path arriving at the mobile at the angle θ has a delay of $\tau_\theta(t)$ and a time-invariant gain a_θ, and the input/output relationship is given by

$$y(t) = \sum_{i=0}^{K-1} a_{\theta_i} x(t - \tau_{\theta_i}(t)) \tag{2.56}$$

The most general version of the model allows the received power distribution $p(\theta)$ and the antenna gain pattern $\alpha(\theta)$ to be arbitrary functions of the angle θ, but the most common scenario assumes uniform power distribution and isotropic antenna gain pattern, i.e., the amplitudes $a_\theta = a/\sqrt{K}$ for all angles θ. This models the situation when the scatterers are located in a ring around the mobile (Figure 2.14). We scale the amplitude of each path by $\sqrt{K}$ so that the total received energy along all paths is a^2; for large K, the received energy along each path is a small fraction of the total energy.

Suppose the communication bandwidth W is much smaller than the reciprocal of the delay spread. The complex baseband channel can be represented by a single tap at each time:

$$y[m] = h_0[m]x[m] + w[m]. \tag{2.57}$$

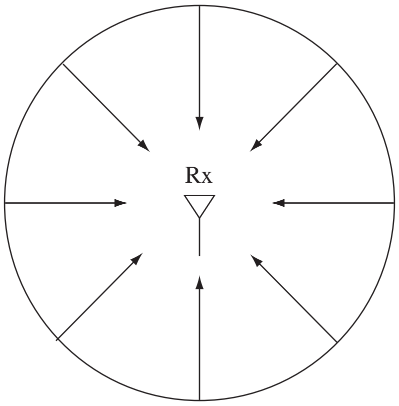


Figure 2.14 The one-ring model.

The phase of the signal arriving at time 0 from an angle θ is $2\pi f_c \tau_\theta(0)$ mod 2π, where f_c is the carrier frequency. Making the assumption that this phase is uniformly distributed in $[0, 2\pi]$ and independently distributed across all angles θ, the tap gain process $\{h_0[m]\}$ is a sum of many small independent contributions, one from each angle. By the Central Limit Theorem, it is reasonable to model the process as Gaussian. Exercise 2.17 shows further that the process is in fact stationary with an autocorrelation function $R_0[n]$ given by:

$$R_0[n] = 2a^2 \pi J_0\left(n\pi D_s/W\right) \tag{2.58}$$

where $J_0(\cdot)$ is the zeroth-order Bessel function of the first kind:

$$J_0(x) := \frac{1}{\pi}\int_0^{\pi} e^{jx\cos\theta} d\theta. \tag{2.59}$$

and $D_s = 2f_c v/c$ is the Doppler spread. The power spectral density $S(f)$, defined on $[-1/2, +1/2]$, is given by

$$S(f) = \begin{cases} \frac{4a^2 W}{D_s\sqrt{1-(2fW/D_s)^2}} & -D_s/(2W) \leqslant f \leqslant +D_s/(2W) \\ 0 & \text{else.} \end{cases} \tag{2.60}$$

This can be verified by computing the inverse Fourier transform of (2.60) to be (2.58). Plots of the autocorrelation function and the spectrum for are shown in Figure 2.15. If we define the coherence time T_c to be the value of n/W such that $R_0[n] = 0.05R_0[0]$, then

$$T_c = \frac{J_0^{-1}(0.05)}{\pi D_s}, \tag{2.61}$$

i.e., the coherence time is inversely proportional to D_s.

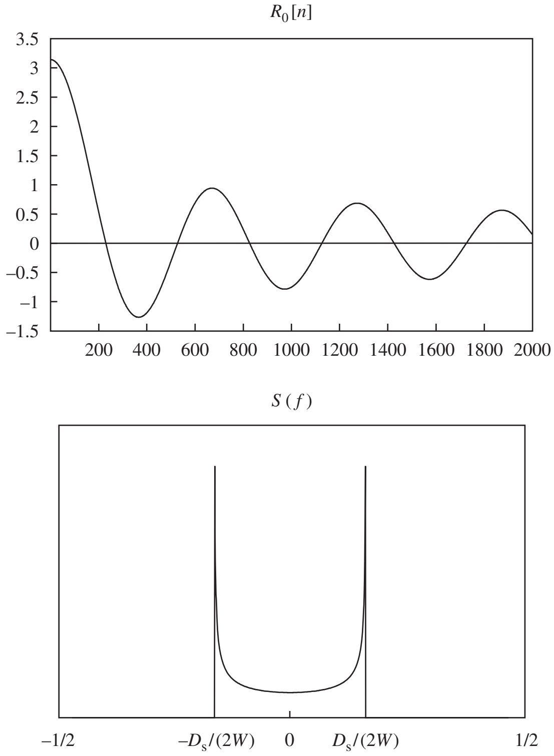


Figure 2.15 Plots of the auto-correlation function and Doppler spectrum in Clarke's model.

In Exercise 2.17, you will also verify that $S(f)\mathrm{d}f$ has the physical interpretation of the received power along paths that have Doppler shifts in the range $[f, f+\mathrm{d}f]$. Thus, $S(f)$ is also called the *Doppler spectrum*. Note that $S(f)$ is zero beyond the maximum Doppler shift.

Chapter 2 The main plot

Large-scale fading

Variation of signal strength over distances of the order of cell sizes. Received power decreases with distance r like:

$$\frac{1}{r^2} \quad \text{(free space)}$$

$$\frac{1}{r^4} \quad \text{(reflection from ground plane).}$$

Decay can be even faster due to shadowing and scattering effects.

Small-scale fading

Variation of signal strength over distances of the order of the carrier wavelength, due to constructive and destructive interference of multipaths. Key parameters:

$$\text{Doppler spread } D_\mathrm{s} \longleftrightarrow \text{coherence time } T_\mathrm{c} \sim 1/D_\mathrm{s}$$

Doppler spread is proportional to the velocity of the mobile and to the angular spread of the arriving paths.

$$\text{delay spread } T_\mathrm{d} \longleftrightarrow \text{coherence bandwidth } W_\mathrm{c} \sim 1/T_\mathrm{d}$$

Delay spread is proportional to the difference between the lengths of the shortest and the longest paths.

Input/output channel models

- Continuous-time passband (2.14):

$$y(t) = \sum_i a_i(t) x(t - \tau_i(t)).$$

- Continuous-time complex baseband (2.26):

$$y_\mathrm{b}(t) = \sum_i a_i(t) \mathrm{e}^{-\mathrm{j}2\pi f_\mathrm{c} \tau_i(t)} x_\mathrm{b}(t - \tau_i(t)).$$

- Discrete-time complex baseband with AWGN (2.38):

$$y[m] = \sum_\ell h_\ell[m] x[m - \ell] + w[m].$$

The ℓth tap is the aggregation of the physical paths with delays in $[\ell/W - 1/(2W), \ell/W + 1/(2W)]$.

Statistical channel models

- $\{h_\ell[m]\}_m$ is modeled as circular symmetric processes independent across the taps.
- If for all taps,

$$h_\ell[m] \sim \mathcal{CN}(0, \sigma_\ell^2),$$

  the model is called *Rayleigh*.
- If for one tap,

$$h_\ell[m] = \sqrt{\frac{\kappa}{\kappa + 1}} \sigma_\ell \mathrm{e}^{\mathrm{j}\theta} + \sqrt{\frac{1}{\kappa + 1}} \mathcal{CN}(0, \sigma_\ell^2),$$

  the model is called *Rician* with K-factor κ.

- The tap gain auto-correlation function $R_\ell[n] := \mathbb{E}[h_\ell^*[0]h_\ell[n]]$ models the dependency over time.
- The delay spread is $1/W$ times the range of taps ℓ which contains most of the total gain $\sum_{\ell=0}^{\infty} R_\ell[0]$. The coherence time is $1/W$ times the range of n for which $R_\ell[n]$ is significantly different from $R_\ell[0]$.

2.5 Bibliographical notes

This chapter was modified from R. G. Gallager's MIT 6.450 course notes on digital communication. The focus is on small-scale multipath fading. Large-scale fading models are discussed in many texts; see for example Rappaport [98]. Clarke's model was introduced in [22] and elaborated further in [62]. Our derivation here of the Clarke power spectrum follows the approach of [111].

2.6 Exercises

Exercise 2.1 (Gallager) Consider the electric field in (2.4).

1. It has been derived under the assumption that the motion is in the direction of the line-of-sight from sending antenna to receive antenna. Find the electric field assuming that ϕ is the angle between the line-of-sight and the direction of motion of the receiver. Assume that the range of time of interest is small enough so that changes in (θ, ψ) can be ignored.
2. Explain why, and under what conditions, it is a reasonable approximation to ignore the change in (θ, ψ) over small intervals of time.

Exercise 2.2 (Gallager) Equation (2.13) was derived under the assumption that $r(t) \approx d$. Derive an expression for the received waveform for general $r(t)$. Break the first term in (2.11) into two terms, one with the same numerator but the denominator $2d - r_0 - vt$ and the other with the remainder. Interpret your result.

Exercise 2.3 In the two-path example in Sections 2.1.3 and 2.1.4, the wall is on the right side of the receiver so that the reflected wave and the direct wave travel in opposite directions. Suppose now that the reflecting wall is on the left side of transmitter. Redo the analysis. What is the nature of the multipath fading, both over time and over frequency? Explain any similarity or difference with the case considered in Sections 2.1.3 and 2.1.4.

Exercise 2.4 A mobile receiver is moving at a speed v and is receiving signals arriving along two reflected paths which make angles θ_1 and θ_2 with the direction of motion. The transmitted signal is a sinusoid at frequency f.

1. Is the above information enough for estimating (i) the coherence time T_c; (ii) the coherence bandwidth W_c? If so, express them in terms of the given parameters. If not, specify what additional information would be needed.
2. Consider an environment in which there are reflectors and scatterers in all directions from the receiver and an environment in which they are clustered within a small

angular range. Using part (1), explain how the channel would differ in these two environments.

Exercise 2.5 Consider the propagation model in Section 2.1.5 where there is a reflected path from the ground plane.

1. Let r_1 be the length of the direct path in Figure 2.6. Let r_2 be the length of the reflected path (summing the path length from the transmitter to the ground plane and the path length from the ground plane to the receiver). Show that $r_2 - r_1$ is asymptotically equal to b/r and find the value of the constant b. *Hint*: Recall that for x small, $\sqrt{1+x} \approx 1 + x/2$ in the sense that $(\sqrt{1+x} - 1)/x \to 1/2$ as $x \to 0$.
2. Assume that the received waveform at the receive antenna is given by

$$E_{\mathrm{r}}(f, t) = \frac{\alpha \cos 2\pi[ft - fr_1/c]}{r_1} - \frac{\alpha \cos 2\pi[ft - fr_2/c]}{r_2}. \tag{2.62}$$

   Approximate the denominator r_2 by r_1 in (2.62) and show that $E_{\mathrm{r}} \approx \beta/r^2$ for r^{-1} much smaller than c/f. Find the value of β.
3. Explain why this asymptotic expression remains valid without first approximating the denominator r_2 in (2.62) by r_1.

Exercise 2.6 Consider the following simple physical model in just a *single* dimension. The source is at the origin and transmits an isotropic wave of angular frequency ω. The physical environment is filled with uniformly randomly located obstacles. We will model the inter-obstacle distance as an exponential random variable, i.e., it has the density[10]

$$\eta \mathrm{e}^{-\eta r}, \qquad r \geq 0. \tag{2.63}$$

Here $1/\eta$ is the mean distance between obstacles and captures the *density* of the obstacles. Viewing the source as a stream of photons, suppose each obstacle independently (from one photon to the other and independent of the behavior of the other obstacles) either absorbs the photon with probability γ or scatters it either to the left or to the right (both with equal probability $(1-\gamma)/2$).

Now consider the path of a photon transmitted either to the left or to the right with equal probability from some fixed point on the line. The probability density function of the distance (denoted by r) to the first obstacle (the distance can be on either side of the starting point, so r takes values on the entire line) is equal to

$$q(r) := \frac{\eta \mathrm{e}^{-\eta |r|}}{2}, \qquad r \in \mathcal{R}. \tag{2.64}$$

So the probability density function of the distance at which the photon is absorbed upon hitting the first obstacle is equal to

$$f_1(r) := \gamma q(r), \qquad r \in \mathcal{R}. \tag{2.65}$$

[10] This random arrangement of points on a line is called a *Poisson point process*.

1. Show that the probability density function of the distance from the origin at which the second obstacle is met is

$$f_2(r) := \int_{-\infty}^{\infty} (1-\gamma)q(x)f_1(r-x)\mathrm{d}x, \qquad r \in \mathcal{R}. \tag{2.66}$$

2. Denote by $f_k(r)$ the probability density function of the distance from the origin at which the photon is absorbed by exactly the kth obstacle it hits and show the recursive relation

$$f_{k+1}(r) = \int_{-\infty}^{\infty} (1-\gamma)q(x)f_k(r-x)\mathrm{d}x, \qquad r \in \mathcal{R}. \tag{2.67}$$

3. Conclude from the previous step that the probability density function of the distance from the source at which the photon is absorbed (by some obstacle), denoted by $f(r)$, satisfies the recursive relation

$$f(r) = \gamma q(r) + (1-\gamma)\int_{-\infty}^{\infty} q(x)f(r-x)\mathrm{d}x, \qquad r \in \mathcal{R}. \tag{2.68}$$

   Hint: Observe that $f(r) = \sum_{k=1}^{\infty} f_k(r)$.
4. Show that

$$f(r) = \frac{\sqrt{\gamma}\eta}{2}\mathrm{e}^{-\eta\sqrt{\gamma}|r|} \tag{2.69}$$

   is a solution to the recursive relation in (2.68). *Hint*: Observe that the convolution between the probability densities $q(\cdot)$ and $f(\cdot)$ in (2.68) is more easily represented using Fourier transforms.
5. Now consider the photons that are absorbed at a distance of more than r from the source. This is the radiated power density at a distance r and is found by integrating $f(x)$ over the range (r, ∞) if $r > 0$ and $(-\infty, r)$ if $r < 0$. Calculate the radiated power density to be

$$\frac{\mathrm{e}^{-\gamma\sqrt{\eta}|r|}}{2}, \tag{2.70}$$

   and conclude that the power decreases exponentially with distance r. Also observe that with very low absorption ($\gamma \to 0$) or very few obstacles ($\eta \to 0$), the power density converges to 0.5; this is expected since the power splits equally on either side of the line.

Exercise 2.7 In Exercise 2.6, we considered a single-dimensional physical model of a scattering and absorption environment and concluded that power decays exponentially with distance. A reading exercise is to study [42], which considers a natural extension of this simple model to two- and three-dimensional spaces. Further, it extends the analysis to two- and three-dimensional physical models. While the analysis is more complicated, we arrive at the same conclusion: the radiated power decays exponentially with distance.

Exercise 2.8 (Gallager) Assume that a communication channel first filters the transmitted passband signal before adding WGN. Suppose the channel is known and the channel filter has an impulse response $h(t)$. Suppose that a QAM scheme with symbol duration T is developed without knowledge of the channel filtering. A baseband filter $\theta(t)$ is developed satisfying the Nyquist property that $\{\theta(t-kT)\}_k$ is an orthonormal set. The matched filter $\theta(-t)$ is used at the receiver before sampling and detection.

If one is aware of the channel filter $h(t)$, one may want to redesign either the baseband filter at the transmitter or the baseband filter at the receiver so that there is no intersymbol interference between receiver samples and so that the noise on the samples is i.i.d.

1. Which filter should one redesign?
2. Give an expression for the impulse response of the redesigned filter (assume a carrier frequency f_c).
3. Draw a figure of the various filters at passband to show why your solution is correct. (We suggest you do this before answering the first two parts.)

Exercise 2.9 Consider the two-path example in Section 2.1.4 with $d = 2\,\text{km}$ and the receiver at 1.5 km from the transmitter moving at velocity 60 km/h away from the transmitter. The carrier frequency is 900 MHz.

1. Plot in MATLAB the magnitudes of the taps of the discrete-time baseband channel at a fixed time t. Give a few plots for several bandwidths W so as to exhibit both flat and frequency-selective fading.
2. Plot the time variation of the phase and magnitude of a typical tap of the discrete-time baseband channel for a bandwidth where the channel is (approximately) flat and for a bandwidth where the channel is frequency-selective. How do the time-variations depend on the bandwidth? Explain.

Exercise 2.10 For each tap of the discrete-time channel response, the Doppler spread is the range of Doppler shifts of the paths contributing to that tap. Give an example of an environment (i.e. location of reflectors/scatterers with respect to the location of the transmitter and the receiver) in which the Doppler spread is the same for different taps and an environment in which they are different.

Exercise 2.11 Verify (2.39) and (2.40).

Exercise 2.12 In this problem we consider generating passband orthogonal waveforms from baseband ones.

1. Show that if the waveforms $\{\theta(t-nT)\}_n$ form an orthogonal set, then the waveforms $\{\psi_{n,1}, \psi_{n,2}\}_n$ also form an orthogonal set, provided that $\theta(t)$ is band-limited to $[-f_c, f_c]$. Here,

$$\begin{aligned}\psi_{n,1}(t) &= \theta(t-nT)\cos 2\pi f_c t,\\ \psi_{n,2}(t) &= \theta(t-nT)\sin 2\pi f_c t.\end{aligned}$$

   How should we normalize the energy of $\theta(t)$ to make the $\psi(t)$ *orthonormal*?
2. For a given f_c, find an example where the result in part (1) is false when the condition that $\theta(t)$ is band-limited to $[-f_c, f_c]$ is violated.

Exercise 2.13 Verify (2.25). Does this equation contain any more information about the communication system in Figure 2.9 beyond what is in (2.24)? Explain.

Exercise 2.14 Compute the probability density function of the magnitude $|X|$ of a complex circular symmetric Gaussian random variable X with variance σ^2.

Exercise 2.15 In the text we have discussed the various reasons why the channel tap gains, $h_\ell[m]$, vary in time (as a function of m) and how the various dynamics operate at different time-scales. The analysis is based on the assumption that communication takes place on a bandwidth W around a carrier frequency f_{c} with $f_{\mathrm{c}} \gg W$. This assumption is not valid for *ultra-wideband* (UWB) communication systems, where the transmission bandwidth is from 3.1 GHz to 10.6 GHz, as regulated by the FCC. Redo the analysis for this system. What is the main mechanism that causes the tap gains to vary at the fastest time-scale, and what is this fastest time-scale determined by?

Exercise 2.16 In Section 2.4.2, we argue that the channel gain $h_\ell[m]$ at a particular time m can be assumed to be circular symmetric. Extend the argument to show that it is also reasonable to assume that the complex random vector

$$\mathbf{h} := \begin{pmatrix} h_\ell[m] \\ h_\ell[m+1] \\ \vdots \\ h_\ell[m+n] \end{pmatrix}$$

is circular symmetric for any n.

Exercise 2.17 In this question, we will analyze in detail Clarke's one-ring model discussed at the end of the chapter. Recall that the scatterers are assumed to be located in a ring around the receiver moving at speed v. There are K paths coming in at angles $\theta_i = 2\pi i/K$ with respect to the direction of motion of the mobile, $i = 0, \ldots, K-1$. The path coming at angle θ has a delay of $\tau_\theta(t)$ and a time-invariant gain $a/\sqrt{K}$ (not dependent on the angle), and the input/output relationship is given by

$$y(t) = \frac{a}{\sqrt{K}} \sum_{i=0}^{K-1} x(t - \tau_{\theta_i}(t)). \tag{2.71}$$

1. Give an expression for the impulse response $h(\tau, t)$ for this channel, and give an expression for $\tau_\theta(t)$ in terms of $\tau_\theta(0)$. (You can assume that the distance the mobile travelled in $[0, t]$ is small compared to the radius of the ring.)
2. Suppose communication takes place at carrier frequency f_c and over a narrowband of bandwidth W such that the delay spread of the channel T_d satisfies $T_d \ll 1/W$. Argue that the discrete-time baseband model can be approximately represented by a single tap

$$y[m] = h_0[m]x[m] + w[m], \tag{2.72}$$

   and give an approximate expression for that tap in terms of the a_θ's and $\tau_\theta(t)$'s. *Hint*: Your answer should contain no sinc functions.
3. Argue that it is reasonable to assume that the phase of the path from an angle θ at time 0,

$$2\pi f_c \tau_\theta(0) \mod 2\pi$$

   is uniformly distributed in $[0, 2\pi]$ and that it is i.i.d. across θ.

4. Based on the assumptions in part (3), for large K one can use the Central Limit Theorem to approximate $\{h_0[m]\}$ as a Gaussian process. Verify that the limiting process is stationary and the autocorrelation function $R_0[n]$ is given by (2.58).
5. Verify that the Doppler spectrum $S(f)$ is given by (2.60). *Hint*: It is easier to show that the inverse Fourier transform of (2.60) is (2.58).
6. Verify that $S(f)\mathrm{d}f$ is indeed the received power from the paths that have Doppler shifts in $[f, f+\mathrm{d}f]$. Is this surprising?

Exercise 2.18 Consider a one-ring model where there are K scatterers located at angles $\theta_i = 2\pi i/K$, $i = 0, \ldots, K-1$, on a circle of radius 1 km around the receiver and the transmitter is 2 km away. (The angles are with respect to the line joining the transmitter and the receiver.) The transmit power is P. The power attenuation along a path from the transmitter to a scatterer to the receiver is

$$\frac{G}{K} \cdot \frac{1}{s^2} \cdot \frac{1}{r^2}, \tag{2.73}$$

where G is a constant and r and s are the distance from the transmitter to the scatterer and the distance from the scatterer to the receiver respectively. Communication takes place at a carrier frequency $f_c = 1.9$ GHz and the bandwidth is W Hz. You can assume that, at any time, the phases of each arriving path in the baseband representation of the channel are independent and uniformly distributed between 0 and 2π.

1. What are the key differences and the similarities between this model and the Clarke's model in the text?
2. Find approximate conditions on the bandwidth W for which one gets a flat fading channel.
3. Suppose the bandwidth is such that the channel is frequency selective. For large K, find approximately the amount of power in tap ℓ of the discrete-time baseband impulse response of the channel (i.e., compute the power-delay profile.). Make any simplifying assumptions but state them. (You can leave your answers in terms of integrals if you cannot evaluate them.)
4. Compute and sketch the power-delay profile as the bandwidth becomes very large (and K is large).
5. Suppose now the receiver is moving at speed v towards the (fixed) transmitter. What is the Doppler spread of tap ℓ? Argue heuristically from physical considerations what the Doppler spectrum (i.e., power spectral density) of tap ℓ is, for large K.
6. We have made the assumptions that the scatterers are all on a circle of radius 1km around the receiver and the paths arrive with independent and uniform distributed phases at the receiver. Mathematically, are the two assumptions consistent? If not, do you think it matters, in terms of the validity of your answers to the earlier parts of this question?

Exercise 2.19 Often in modeling multiple input multiple output (MIMO) fading channels the fading coefficients between different transmit and receive antennas are assumed to be independent random variables. This problem explores whether this is a reasonable assumption based on Clarke's one-ring scattering model and the antenna separation.

1. (Antenna separation at the mobile) Assume a mobile with velocity v moving away from the base-station, with uniform scattering from the ring around it.

(a) Compute the Doppler spread D_s for a carrier frequency f_c, and the corresponding coherence time T_c.

(b) Assuming that fading states separated by T_c are approximately uncorrelated, at what distance should we place a second antenna at the mobile to get an independently faded signal? *Hint*: How much distance does the mobile travel in T_c?

2. (Antenna separation at the base-station) Assume that the scattering ring has radius R and that the distance between the base-station and the mobile is d. Further assume for the time being that the base-station is moving away from the mobile with velocity v'. Repeat the previous part to find the minimum antenna spacing at the base-station for uncorrelated fading. *Hint*: Is the scattering still uniform around the base-station?
3. Typically, the scatterers are local around the mobile (near the ground) and far away from the base-station (high on a tower). What is the implication of your result in part (2) for this scenario?

CHAPTER

3 Point-to-point communication: detection, diversity, and channel uncertainty

In this chapter we look at various basic issues that arise in communication over fading channels. We start by analyzing uncoded transmission in a narrowband fading channel. We study both coherent and non-coherent detection. In both cases the error probability is much higher than in a non-faded AWGN channel. The reason is that there is a significant probability that the channel is in a deep fade. This motivates us to investigate various *diversity* techniques that improve the performance. The diversity techniques operate over time, frequency or space, but the basic idea is the same. By sending signals that carry the same information through different paths, multiple independently faded replicas of data symbols are obtained at the receiver end and more reliable detection can be achieved. The simplest diversity schemes use *repetition coding*. More sophisticated schemes exploit channel diversity and, at the same time, efficiently use the degrees of freedom in the channel. Compared to repetition coding, they provide *coding gains* in addition to *diversity gains*. In space diversity, we look at both transmit and receive diversity schemes. In frequency diversity, we look at three approaches:

- single-carrier with inter-symbol interference equalization,
- direct-sequence spread-spectrum,
- orthogonal frequency division multiplexing.

Finally, we study the impact of channel uncertainty on the performance of diversity combining schemes. We will see that, in some cases, having too many diversity paths can have an adverse effect due to channel uncertainty.

To familiarize ourselves with the basic issues, the emphasis of this chapter is on concrete techniques for communication over fading channels. In Chapter 5 we take a more fundamental and systematic look and use information theory to derive the *best* performance one can achieve. At that fundamental level, we will see many of the issues discussed here recur.

The derivations in this chapter make repeated use of a few key results in vector detection under Gaussian noise. We develop and summarize the basic results in Appendix A, emphasizing the underlying geometry. The reader is encouraged to take a look at the appendix before proceeding with this chapter and to refer back to it often. In particular, a thorough understanding of the canonical detection problem in Summary A.2 will be very useful.

3.1 Detection in a Rayleigh fading channel

3.1.1 Non-coherent detection

We start with a very simple detection problem in a fading channel. For simplicity, let us assume a flat fading model where the channel can be represented by a single discrete-time complex filter tap $h_0[m]$, which we abbreviate as $h[m]$:

$$y[m] = h[m]x[m] + w[m], \tag{3.1}$$

where $w[m] \sim \mathcal{CN}(0, N_0)$. We suppose Rayleigh fading, i.e., $h[m] \sim \mathcal{CN}(0, 1)$, where we normalize the variance to be 1. For the time being, however, we do not specify the dependence between the fading coefficients $h[m]$ at different times m nor do we make any assumption on the prior knowledge the receiver might have of $h[m]$. (This latter assumption is sometimes called *non-coherent* communication.)

First consider uncoded binary antipodal signaling (or binary phase-shift-keying, BPSK) with amplitude a, i.e., $x[m] = \pm a$, and the symbols $x[m]$ are independent over time. This signaling scheme fails completely, even in the absence of noise, since the phase of the received signal $y[m]$ is uniformly distributed between 0 and 2π regardless of whether $x[m] = a$ or $x[m] = -a$ is transmitted. Further, the received amplitude is independent of the transmitted symbol. Binary antipodal signaling is binary phase modulation and it is easy to see that phase modulation in general is similarly flawed. Thus, signal structures are required in which either different signals have different magnitudes, or coding between symbols is used. Next we look at orthogonal signaling, a special type of coding between symbols.

Consider the following simple orthogonal modulation scheme: a form of binary pulse-position modulation. For a pair of time samples, transmit either

$$\mathbf{x}_A := \begin{pmatrix} x[0] \\ x[1] \end{pmatrix} = \begin{pmatrix} a \\ 0 \end{pmatrix}, \tag{3.2}$$

or

$$\mathbf{x}_B := \begin{pmatrix} 0 \\ a \end{pmatrix}. \tag{3.3}$$

We would like to perform detection based on

$$\mathbf{y} := \begin{pmatrix} y[0] \\ y[1] \end{pmatrix}. \tag{3.4}$$

This is a simple hypothesis testing problem, and it is straightforward to derive the maximum likelihood (ML) rule:

$$\Lambda(\mathbf{y}) \underset{\mathbf{x}_B}{\overset{\mathbf{x}_A}{\gtrless}} 0, \tag{3.5}$$

where $\Lambda(\mathbf{y})$ is the log-likelihood ratio

$$\Lambda(\mathbf{y}) := \ln\left\{\frac{f(\mathbf{y}|\mathbf{x}_A)}{f(\mathbf{y}|\mathbf{x}_B)}\right\}. \tag{3.6}$$

It can be seen that, if $\mathbf{x}_A$ is transmitted, $y[0] \sim \mathcal{CN}(0, a^2+N_0)$ and $y[1] \sim \mathcal{CN}(0, N_0)$ and $y[0], y[1]$ are independent. Similarly, if $\mathbf{x}_B$ is transmitted, $y[0] \sim \mathcal{CN}(0, N_0)$ and $y[1] \sim \mathcal{CN}(0, a^2+N_0)$. Further, $y[0]$ and $y[1]$ are independent. Hence the log-likelihood ratio can be computed to be

$$\Lambda(\mathbf{y}) = \frac{\left\{|y[0]|^2 - |y[1]|^2\right\} a^2}{(a^2+N_0)N_0}. \tag{3.7}$$

The optimal rule is simply to decide $\mathbf{x}_A$ is transmitted if $|y[0]|^2 > |y[1]|^2$ and decide $\mathbf{x}_B$ otherwise. Note that the rule does not make use of the phases of the received signal, since the random unknown phases of the channel gains $h[0], h[1]$ render them useless for detection. Geometrically, we can interpret the detector as projecting the received vector $\mathbf{y}$ onto each of the two possible transmit vectors $\mathbf{x}_A$ and $\mathbf{x}_B$ and comparing the energies of the projections (Figure 3.1). Thus, this detector is also called an *energy* or a *square-law* detector. It is somewhat surprising that the optimal detector does not depend on how $h[0]$ and $h[1]$ are correlated.

We can analyze the error probability of this detector. By symmetry, we can assume that $\mathbf{x}_A$ is transmitted. Under this hypothesis, $y[0]$ and $y[1]$ are

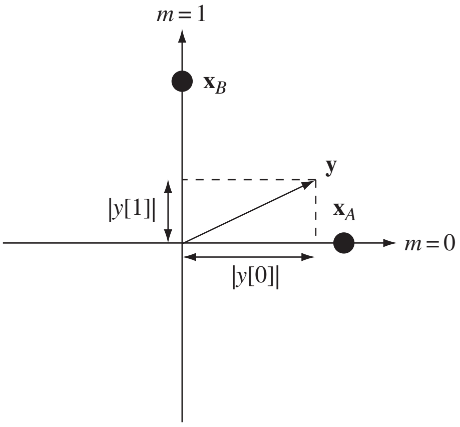


Figure 3.1 The non-coherent detector projects the received vector **y** onto each of the two orthogonal transmitted vectors $\mathbf{x}_A$ and $\mathbf{x}_B$ and compares the lengths of the projections.

independent circular symmetric complex Gaussian random variables with variances a^2+N_0 and N_0 respectively. (See Section A.1.3 in the appendices for a discussion on circular symmetric Gaussian random variables and vectors.) As shown there, $|y[0]|^2$, $|y[1]|^2$ are exponentially distributed with mean a^2+N_0 and N_0 respectively.[1] The probability of error can now be computed by direct integration:

$$p_{\mathrm{e}} = \mathbb{P}\left\{|y[1]|^2 > |y[0]|^2 | \mathbf{x}_A\right\} = \left[2+\frac{a^2}{N_0}\right]^{-1}. \tag{3.8}$$

We make the general definition

$$\boxed{\mathsf{SNR} := \frac{\text{average received signal energy per (complex) symbol time}}{\text{noise energy per (complex) symbol time}}} \tag{3.9}$$

which we use consistently throughout the book *for any modulation scheme*. The noise energy per complex symbol time is N_0.[2] For the orthogonal modulation scheme here, the average received energy per symbol time is $a^2/2$ and so

$$\mathsf{SNR} := \frac{a^2}{2N_0}. \tag{3.10}$$

Substituting into (3.8), we can express the error probability of the orthogonal scheme in terms of SNR:

$$p_{\mathrm{e}} = \frac{1}{2(1+\mathsf{SNR})}. \tag{3.11}$$

This is a very discouraging result. To get an error probability $p_{\mathrm{e}} = 10^{-3}$ one would require $\mathsf{SNR} \approx 500$ (27 dB). Stupendous amounts of power would be required for more reliable communication.

3.1.2 Coherent detection

Why is the performance of the non-coherent maximum likelihood (ML) receiver on a fading channel so bad? It is instructive to compare its performance with detection in an AWGN channel without fading:

$$y[m] = x[m] + w[m]. \tag{3.12}$$

[1] Recall that a random variable U is exponentially distributed with mean μ if its pdf is $f_U(u) = \frac{1}{\mu}\mathrm{e}^{-u/\mu}$.

[2] The orthogonal modulation scheme considered here uses only real symbols and hence transmits only on the I channel. Hence it may seem more natural to define the SNR in terms of noise energy per *real* symbol, i.e., $N_0/2$. However, later we will consider modulation schemes that use complex symbols and hence transmit on both the I and Q channels. In order to be consistent throughout, we choose to define SNR this way.

For antipodal signaling (BPSK), $x[m] = \pm a$, a sufficient statistic is $\Re\{y[m]\}$ and the error probability is

$$p_\mathrm{e} = Q\left(\frac{a}{\sqrt{N_0/2}}\right) = Q\left(\sqrt{2\mathsf{SNR}}\right), \tag{3.13}$$

where $\mathsf{SNR} = a^2/N_0$ is the received signal-to-noise ratio per symbol time, and $Q(\cdot)$ is the complementary cumulative distribution function of an $N(0, 1)$ random variable. This function decays exponentially with x^2; more specifically,

$$Q(x) < \mathrm{e}^{-x^2/2}, \qquad x > 0 \tag{3.14}$$

and

$$Q(x) > \frac{1}{\sqrt{2\pi}x}\left(1 - \frac{1}{x^2}\right)\mathrm{e}^{-x^2/2}, \qquad x > 1. \tag{3.15}$$

Thus, *the detection error probability decays exponentially in SNR in the AWGN channel while it decays only inversely with the SNR in the fading channel*. To get an error probability of 10^{-3}, an SNR of only about 7 dB is needed in an AWGN channel (as compared to 27 dB in the non-coherent fading channel). Note that $2\sqrt{\mathsf{SNR}}$ is the separation between the two constellation points as a multiple of the standard deviation of the Gaussian noise; the above observation says that when this separation is much larger than 1, the error probability is very small.

Compared to detection in the AWGN channel, the detection problem considered in the previous section has two differences: the channel gains $h[m]$ are random, and the receiver is assumed not to know them. Suppose now that the channel gains are tracked at the receiver so that they are known at the receiver (but still random). In practice, this is done either by sending a known sequence (called a *pilot* or training sequence) or in a decision directed manner, estimating the channel using symbols detected earlier. The accuracy of the tracking depends, of course, on how fast the channel varies. For example, in a narrowband 30-kHz channel (such as that used in the North American TDMA cellular standard IS-136) with a Doppler spread of 100 Hz, the coherence time T_c is roughly 80 symbols and in this case the channel can be estimated with minimal overhead expended in the pilot.[3] For our current purpose, let us suppose that the channel estimates are perfect.

Knowing the channel gains, *coherent* detection of BPSK can now be performed on a symbol by symbol basis. We can focus on one symbol time and drop the time index

$$y = hx + w \tag{3.16}$$

[3] The channel estimation problem for a broadband channel with many taps in the impulse response is more difficult; we will get to this in Section 3.5.

Detection of x from y can be done in a way similar to that in the AWGN case; the decision is now based on the sign of the real sufficient statistic

$$r := \Re\{(h/|h|)^* y\} = |h|x + z, \tag{3.17}$$

where $z \sim N(0, N_0/2)$. If the transmitted symbol is $x = \pm a$, then, for a given value of h, the error probability of detecting x is

$$Q\left(\frac{a|h|}{\sqrt{N_0/2}}\right) = Q\left(\sqrt{2|h|^2 \mathsf{SNR}}\right) \tag{3.18}$$

where $\mathsf{SNR} = a^2/N_0$ is the average received signal-to-noise ratio per symbol time. (Recall that we normalized the channel gain such that $\mathbb{E}[|h|^2] = 1$.) We average over the random gain h to find the overall error probability. For Rayleigh fading when $h \sim \mathcal{CN}(0, 1)$, direct integration yields

$$p_e = \mathbb{E}\left[Q\left(\sqrt{2|h|^2 \mathsf{SNR}}\right)\right] = \frac{1}{2}\left(1 - \sqrt{\frac{\mathsf{SNR}}{1 + \mathsf{SNR}}}\right). \tag{3.19}$$

(See Exercise 3.1.) Figure 3.2 compares the error probabilities of coherent BPSK and non-coherent orthogonal signaling over the Rayleigh fading channel, as well as BPSK over the AWGN channel. We see that while the error probability for BPSK over the AWGN channel decays very fast with the SNR, the error probabilities for the Rayleigh fading channel are much worse,

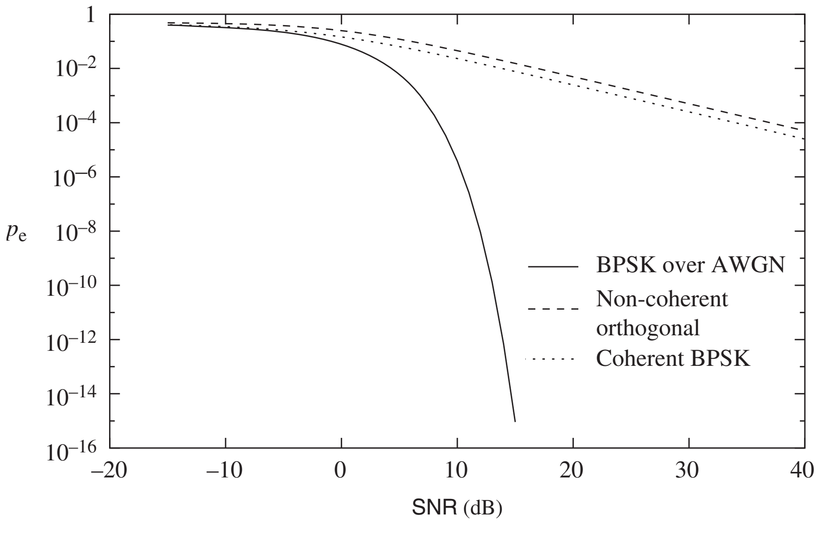


Figure 3.2 Performance of coherent BPSK vs. non-coherent orthogonal signaling over Rayleigh fading channel vs. BPSK over AWGN schannel.

whether the detection is coherent or non-coherent. At high SNR, Taylor series expansion yields

$$\sqrt{\frac{\mathsf{SNR}}{1+\mathsf{SNR}}} = 1 - \frac{1}{2\mathsf{SNR}} + O\left(\frac{1}{\mathsf{SNR}^2}\right). \tag{3.20}$$

Substituting into (3.19), we get the approximation

$$p_{\mathrm{e}} \approx \frac{1}{4\mathsf{SNR}}, \tag{3.21}$$

which decays inversely proportional to the SNR, just as in the non-coherent orthogonal signaling scheme (cf. (3.11)). There is only a 3 dB difference in the required SNR between the coherent and non-coherent schemes; in contrast, at an error probability of 10^{-3}, there is a 17 dB difference between the performance on the AWGN channel and coherent detection on the Rayleigh fading channel.[4]

We see that the main reason why detection in the fading channel has poor performance is not because of the lack of knowledge of the channel at the receiver. It is due to the fact that the channel gain is random and there is a significant probability that the channel is in a "deep fade". At high SNR, we can in fact be more precise about what a "deep fade" means by inspecting (3.18). The quantity $|h|^2\mathsf{SNR}$ is the instantaneous received SNR. Under typical channel conditions, i.e., $|h|^2\mathsf{SNR} \gg 1$, the conditional error probability is very small, since the tail of the Q-function decays very rapidly. In this regime, the separation between the constellation points is much larger than the standard deviation of the Gaussian noise. On the other hand, when $|h|^2\mathsf{SNR}$ is of the order of 1 or less, the separation is of the same order as the standard deviation of the noise and the error probability becomes significant. The probability of this event is

$$\mathbb{P}\{|h|^2\mathsf{SNR} < 1\} = \int_0^{1/\mathsf{SNR}} \mathrm{e}^{-x}\mathrm{d}x \tag{3.22}$$

$$= \frac{1}{\mathsf{SNR}} + O\left(\frac{1}{\mathsf{SNR}^2}\right). \tag{3.23}$$

This probability has the same order of magnitude as the error probability itself (cf. (3.21)). Thus, we can define a "deep fade" via an order-of-magnitude approximation:

$$\boxed{\begin{aligned} \text{Deep fade event} : |h|^2 &< \frac{1}{\mathsf{SNR}}. \\ \mathbb{P}\{\text{deep fade}\} &\approx \frac{1}{\mathsf{SNR}}. \end{aligned}}$$

[4] Communication engineers often compare schemes based on the difference in the required SNR to attain the same error probability. This corresponds to the horizontal gap between the error probability versus SNR curves of the two schemes.

We conclude that high-SNR error events most often occur because the channel is in deep fade and not as a result of the additive noise being large. In contrast, in the AWGN channel the only possible error mechanism is for the additive noise to be large. Thus, the error probability performance over the AWGN channel is much better.

We have used the explicit error probability expression (3.19) to help identify the typical error event at high SNR. We can in fact turn the table around and use it as a basis for an approximate analysis of the high-SNR performance (Exercises 3.2 and 3.3). Even though the error probability p_{e} can be directly computed in this case, the approximate analysis provides much insight as to how typical errors occur. Understanding typical error events in a communication system often suggests how to improve it. Moreover, the approximate analysis gives some hints as to how robust the conclusion is to the Rayleigh fading model. In fact, the only aspect of the Rayleigh fading model that is important to the conclusion is the fact that $\mathbb{P}\{|h|^2 < \epsilon\}$ is proportional to ϵ for ϵ small. This holds whenever the pdf of $|h|^2$ is positive and continuous at 0.

3.1.3 From BPSK to QPSK: exploiting the degrees of freedom

In Section 3.1.2, we have considered BPSK modulation, $x[m] = \pm a$. This uses only the real dimension (the I channel), while in practice both the I and Q channels are used simultaneously in coherent communication, increasing spectral efficiency. Indeed, an extra bit can be transmitted by instead using QPSK (quadrature phase-shift-keying) modulation, i.e., the constellation is

$$\{a(1+j), a(1-j), a(-1+j), a(-1-j)\}; \tag{3.24}$$

in effect, a BPSK symbol is transmitted on each of the I and Q channels simultaneously. Since the noise is independent across the I and Q channels, the bits can be detected separately and the bit error probability on the AWGN channel (cf. (3.12)) is

$$Q\left(\sqrt{\frac{2a^2}{N_0}}\right), \tag{3.25}$$

the same as BPSK (cf. (3.13)). For BPSK, the SNR (as defined in (3.9)) is given by

$$\mathsf{SNR} = \frac{a^2}{N_0}, \tag{3.26}$$

while for QPSK,

$$\mathsf{SNR} = \frac{2a^2}{N_0}, \tag{3.27}$$

is twice that of BPSK since both the I and Q channels are used. Equivalently, for a given SNR, the bit error probability of BPSK is $Q(\sqrt{2\mathsf{SNR}})$ (cf. (3.13)) and that of QPSK is $Q(\sqrt{\mathsf{SNR}})$. The error probability of QPSK under Rayleigh fading can be similarly obtained by replacing SNR by $\mathsf{SNR}/2$ in the corresponding expression (3.19) for BPSK to yield

$$p_{\mathrm{e}} = \frac{1}{2}\left(1 - \sqrt{\frac{\mathsf{SNR}}{2+\mathsf{SNR}}}\right) \approx \frac{1}{2\mathsf{SNR}}. \tag{3.28}$$

at high SNR. For expositional simplicity, we will consider BPSK modulation in many of the discussions in this chapter, but the results can be directly mapped to QPSK modulation.

One important point worth noting is that it is much more energy-efficient to use both the I and Q channels rather than just one of them. For example, if we had to send the two bits carried by the QPSK symbol on the I channel alone, then we would have to transmit a 4-PAM symbol. The constellation is $\{-3b, -b, b, 3b\}$ and the average error probability on the AWGN channel is

$$\frac{3}{2}Q\left(\sqrt{\frac{2b^2}{N_0}}\right). \tag{3.29}$$

To achieve approximately the same error probability as QPSK, the argument inside the Q-function should be the same as that in (3.25) and hence b should be the same as a, i.e., the same minimum separation between points in the two constellations (Figure 3.3). But QPSK requires a transmit energy of $2a^2$ per symbol, while 4-PAM requires a transmit energy of $5b^2$ per symbol. Hence, for the same error probability, approximately 2.5 times more transmit energy is needed: a 4 dB worse performance. Exercise 3.4 shows that this loss is even more significant for larger constellations. The loss is due to the fact that it is more energy efficient to pack, for a desired minimum distance separation, a

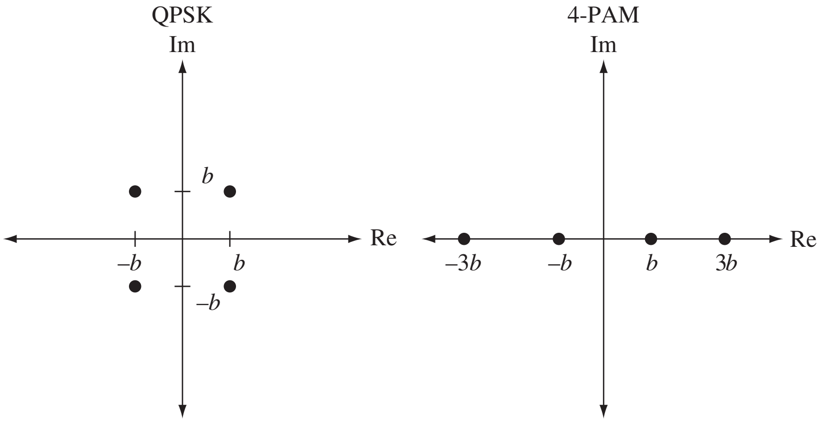


Figure 3.3 QPSK versus 4-PAM: for the same minimum separation between constellation points, the 4-PAM constellation requires higher transmit power.